Monte Carlo methods

Monte Carlo Principle and MCMC

A. Doucet

Carcans

Sept. 2011

Overview of the Lectures

1. Monte Carlo Principles
2. Markov chain Monte Carlo methods.
3. Sequential Monte Carlo methods.
4. Combining MCMC and SMC methods.

A Bayesian motivation

- Bayesian model: Prior $p(\theta)$ and likelihood $f(y|\theta)$

$$\pi(\theta|y) = \frac{p(\theta) f(y|\theta)}{\int_\Theta p(\theta) f(y|\theta) d\theta}.$$

- Except for simple cases -conjugate priors-, there is no closed form expression for the posterior.
- Bayes rule requires one to be able to compute the potentially high dimensional integral

$$\int_\Theta p(\theta) f(y|\theta) d\theta.$$

Point estimates and expectations

- In practice, one is interested in point estimates

$$\begin{aligned}\mathbb{E}\left[\theta \mid y\right] &= \int \theta \pi\left(\theta \mid y\right) d\theta \\ Var\left[\theta \mid y\right] &= \int \theta^2 \pi\left(\theta \mid y\right) d\theta - \mathbb{E}^2\left[\theta \mid y\right]\end{aligned}$$

- But also marginal distributions; e.g. if $\theta = (\theta_1, \theta_2)$ and θ_2 are so-called nuisance parameters then

$$\pi\left(\theta_1 \mid y\right) = \int \pi\left(\theta_1, \theta_2 \mid y\right) d\theta_2 \ .$$

- We might also be interested in

$$\theta_1^{\text{MMAP}} = \arg\max \ \pi\left(\theta_1 \mid y\right)$$

More expectations

- If one is interested in predicting $\tilde{Y} \sim f(\tilde{y}|\theta)$ given y then the predictive density is

$$g(\tilde{y}|y) = \int f(\tilde{y}|\theta)\,\pi(\theta|y)\,d\theta$$

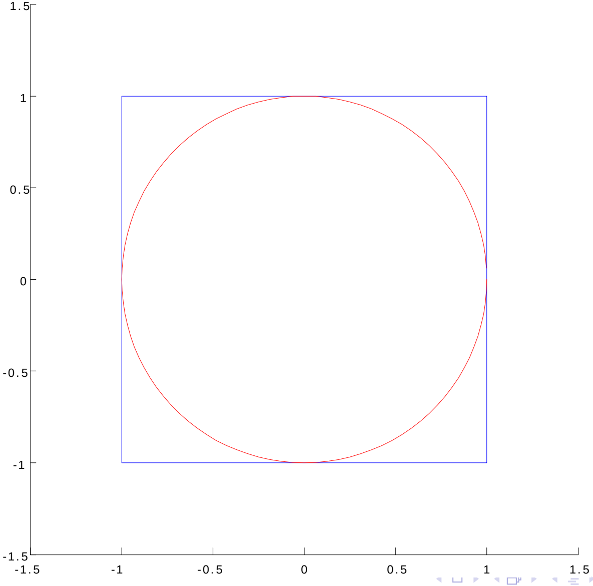
and the corresponding expected value is

$$\mathbb{E}\left[\tilde{Y}|y\right] = \int\int y f(\tilde{y}|\theta)\,\pi(\theta|y)\,d\theta.$$

- For model selection with an infinitely countable number of models

$$\pi(k,\theta_k|\,y) = \frac{\pi(k)\,\pi(\theta_k|k)\,f(y|\,k,\theta_k)}{\sum_{k=1}^{\infty}\pi(k)\int\pi(\theta_k|k)\,f(y|\,k,\theta_k)\,d\theta_k}$$

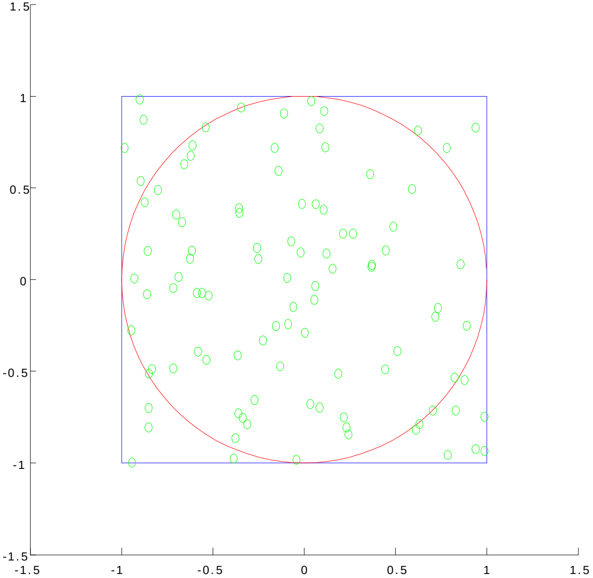
Introduction to Monte Carlo



- An “idealised” rain falls *uniformly* on the square $\mathcal{S}$, *i.e.* the probability for a drop to fall in a region $\mathcal{A}$ is proportional to the area of $\mathcal{A}$.
- Let D be the random variable defined on $\mathsf{X} = \mathcal{S}$ representing the location of a drop and $\mathcal{A}$ a region of the square, then

$$\mathbb{P}(D \in \mathcal{A}) = \frac{\int_{\mathcal{A}} dxdy}{\int_{\mathcal{S}} dxdy} .$$

  where x and y are the Cartesian coordinates.
- Assume we observe N such *independent* drops, say $\{D^{(i)}; i = 1, \ldots, N\}$.


1.5
1
0.5
0
-0.5
-1
-1.5
-1.5
-1
-0.5
0
0.5
1
1.5

A heuristic?

- Intuitively, imagining that you have never followed any statistics course, a sensible technique to estimate the probability $\mathbb{P}(D \in \mathcal{A})$ of falling in a given region $\mathcal{A} \subset \mathcal{S}$ (and think for example of $\mathcal{A} = \mathcal{D}$) would consist of using

$$\mathbb{P}(D \in \mathcal{A}) \approx \frac{\text{number of drops that fell in } \mathcal{A}}{N}.$$

- We want a statistical justification to this.

Probabilities as expectations

- Let us denote the indicator function of a set $\mathcal{A}$ as follows,

$$\mathbb{I}_{\mathcal{A}}(x,y) = \begin{cases} 1 & \text{if point } d = (x,y) \in \mathcal{A}, \\ 0 & \text{otherwise.} \end{cases}$$

- We have

$$\mathbb{P}(D \in \mathcal{A}) = \frac{\int_{\mathcal{S}} \mathbb{I}_{\mathcal{A}}(x,y)\, dxdy}{\int_{\mathcal{S}} dxdy} = \int_{\mathcal{S}} \mathbb{I}_{\mathcal{A}}(x,y) \frac{1}{4} dxdy.$$

- $1/4$ is the probability density associated to $\mathbb{P}$, *i.e.* the density of the uniform distribution on $\mathcal{S}$ denoted $\mathcal{U}_{\mathcal{S}}$.
- Let us define the r.v. $V(D) := \mathbb{I}_{\mathcal{A}}(D) := \mathbb{I}_{\mathcal{A}}(X,Y)$, where X, Y are the rvs representing the Cartesian coordinates of a uniformly distributed point on $\mathcal{S}$ then

$$\mathbb{P}(D \in \mathcal{A}) = \int_{\mathcal{S}} \mathbb{I}_{\mathcal{A}}(x,y) \frac{1}{4} dxdy = \mathbb{E}_{\mathcal{U}_{\mathcal{S}}}(V).$$

Law of large numbers

- Introduce $\{V^{(i)} := V(D^{(i)}), i = 1, \ldots, N\}$ the r.v.s associated to the drops $\{D^{(i)}, i = 1, \ldots, N\}$ and consider the sum

$$S_N = \frac{\sum_{i=1}^N V^{(i)}}{N} = \frac{\text{number of drops that fell in } \mathcal{A}}{N}$$

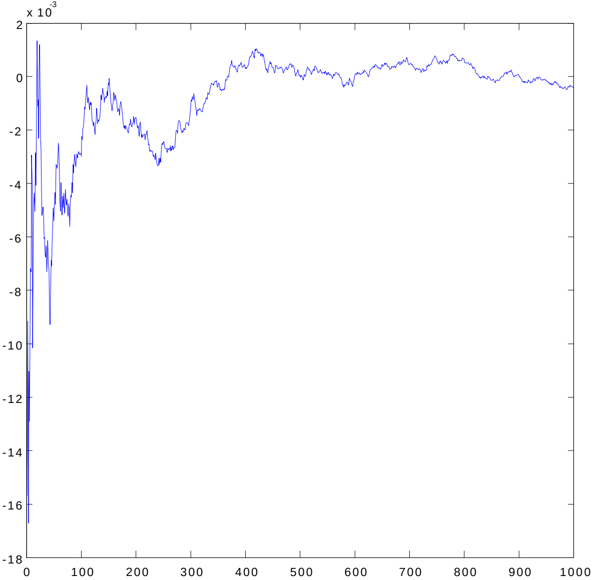
- This expression shows that our suggested approximation of $\mathbb{P}(D \in \mathcal{A})$ is the empirical average of i.i.d. r.v.s $\{V^{(i)}, i = 1, \ldots, N\}$.
- Assuming that the rain lasts forever (i.e. $N \rightarrow +\infty$) then the *law of large numbers* (since $\mathbb{E}_{\mathcal{U}_\mathcal{S}}(|V|) < +\infty$ here) yields

$$\lim_{N \rightarrow +\infty} S_N = \mathbb{E}_{\mathcal{U}_\mathcal{S}}(V), \text{ (almost surely)},$$

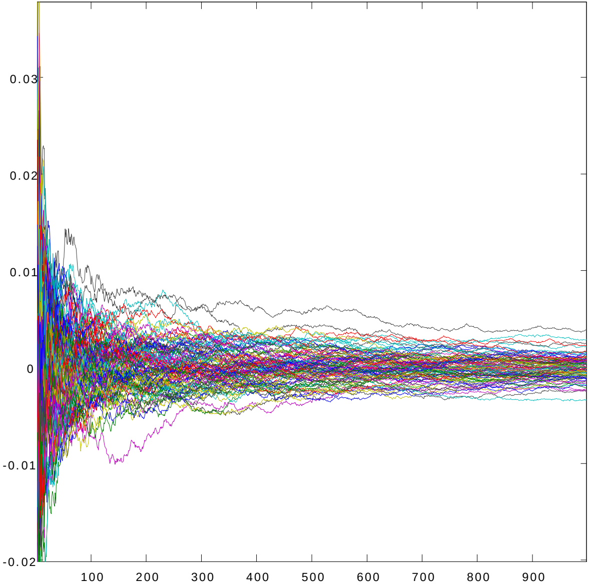
  where we have already proved that $\mathbb{P}(D \in \mathcal{A}) = \mathbb{E}_{\mathcal{U}_\mathcal{S}}(V)$.
- When N is sufficiently large, this mathematically justifies our heuristic method.

Approximating pi

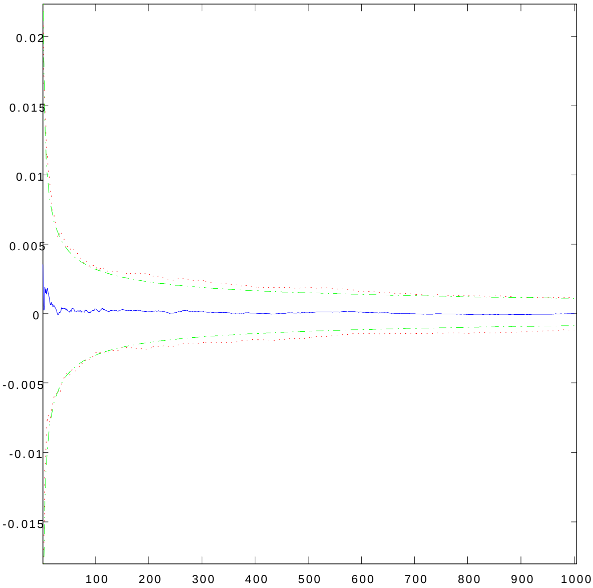
- As we have
$$\mathbb{P}(D \in \mathcal{D}) = \int_{\mathcal{D}} \frac{1}{4} dxdy = \frac{\pi}{4}$$
then S_N is an (unbiased) estimator of $\pi/4$.
- To characterise the precision of our estimator, we can use
$$var(S_N) = \frac{1}{N^2} \sum_{i=1}^{N} var(V^{(i)}) = \frac{1}{N} var(V^{(1)})$$
as the $\{V^{(i)}, i = 1, \ldots, N\}$ are independent.
- One can invoke an asymptotic result, the *central limit theorem* (which can be applied here as $var(V) < +\infty$). As $N \rightarrow +\infty$,
$$\sqrt{N}(S_N - \pi/4) \rightarrow_d \mathcal{N}(0, var(V))$$
which implies that for N large enough the probability of the error being larger than $2\sqrt{var(V)/N}$ (here $2\sqrt{var(V)} = 0.8211$) is
$$\mathbb{P}\left(|S_N - \pi/4| > 2\sqrt{var(V)/N}\right) \simeq 0.05.$$



Convergence of $S_N - \frac{\pi}{4}$ as a function of N for 1 realisation.



Convergence of $S_N - \frac{\pi}{4}$ for 100 realisations.



Square root empirical mean square error $S_N - \frac{\pi}{4}$ accross 100 realisations as a function of N (dashed) and $\pm\sqrt{var(V)/N}$ (dot/dash).

A first generalisation

- Consider the case where $\mathsf{X} = \mathbb{R}^{n_x}$ for any n_x, and in particular $n_x >> 1$. Replace the $\mathcal{S}$ and $\mathcal{D}$ above with a hypercube $\mathcal{S}^{n_x}$ and an inscribed hyperball $\mathcal{D}^{n_x}$ in X.
- Arguments that lead earlier to the formal validation of the Monte Carlo approach remain identical here.
  - In particular the rate of convergence of the estimator in the mean square sense is again in $1/N$ and *independent of the dimension* n_x.
  - This would not be the case using a deterministic method on a grid of regularly spaced points where the CV rate is typically of the form $1/N^{r/n_x}$ where r is related to the smoothness of the contours of $\mathcal{A}$.
  - This is one of the main reasons why Monte Carlo methods are attractive.

- Sometimes it is claimed that Monte Carlo beat the curse of dimensionality....

Well sort of...

- **Example**: Assume you are interested in computing the volume of the hypersphere of radius $R = 1$ in n_x-dimension

$$vol\left(S_{n_x}\right) = \frac{\pi^{\frac{n_x}{2}}}{\Gamma\left(\frac{n_x}{2}+1\right)} \to 0 \text{ as } n_x \to \infty$$

  using samples from the hypercube $[-1,1]^{n_x}$ of volume 2^{n_x}.

- For N samples, the variance of our MC estimate is indeed in

$$\frac{var(X)}{N} = \frac{p_{n_x}\left(1-p_{n_x}\right)}{N} \approx \frac{p_{n_x}}{N} \text{ for large } n_x$$

  where $X \sim$Bernoulli(p_{n_x}) with $p_{n_x} = 2^{-n_x} vol\left(S_{n_x}\right)$.

- We are interested in calculating $2^{n_x}\mathbb{E}(X) = 2^{n_x} p_{n_x}$ but

$$\frac{var(X)}{\mathbb{E}^2(X)} \approx \frac{1}{N.p_{n_x}}$$

- So to get a reasonable relative error, we would need $N \approx 100 p_{n_x}^{-1}$.
- For $n_x = 20$, we have $N \approx 4.06 \times 10^9$ and for $n_x = 40$, $N \approx 3.02 \times 10^{20}$; i.e. if you look for a needle in a haystack, using a blind Monte Carlo strategy will not help much.
- Always keep in mind that what is often important is the relative error you commit, not the absolute error. Even if $\mathbb{E}(X) = 1$, if we have $var(X) = C\alpha^{n_x}$ where $\alpha > 1$ then

$$\frac{var(X)}{N} \leq \varepsilon \Rightarrow N \geq \frac{C\alpha^{n_x}}{\varepsilon};$$

i.e. an number of samples exponential in the dimension n_x is required to ensure a fixed precision.

Further generalisations

- Now we generalise this idea to tackle the generic problem of estimating
$$\mathbb{E}_\pi(f(X)) \triangleq \int_{\mathsf{X}} f(x)\pi(x)dx,$$
where $f : \mathsf{X} \to \mathbb{R}^{n_f}$ and π is a probability distribution on $\mathsf{X} \subset \mathbb{R}^{n_x}$.
- We will assume that $\mathbb{E}_\pi(|f(X)|) < +\infty$ but that it is difficult to obtain an analytical expression for $\mathbb{E}_\pi(f(X))$.
- Here π is any probability density function.

Empirical Measure

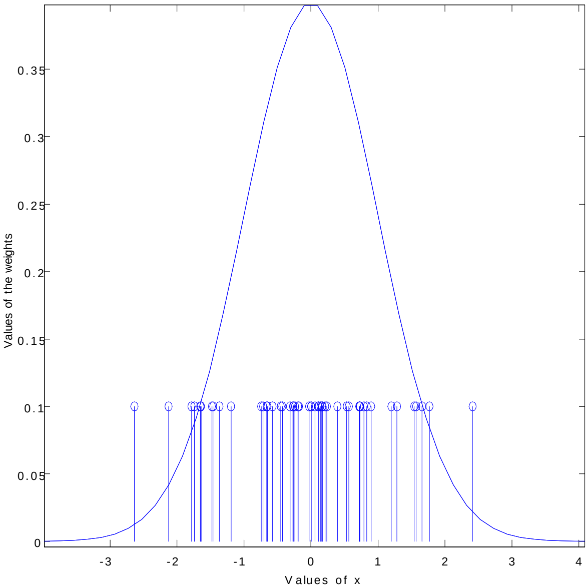
- Let us introduce the delta-Dirac function δ_{x_0} for $x_0 \in \mathsf{X}$ defined for any $f : \mathsf{X} \to \mathbb{R}^{n_f}$ as follows

$$\int_{\mathsf{X}} f(x)\delta_{x_0}(x)dx = f(x_0)$$

- Assume $N >> 1$ *i.i.d.* samples $X^{(i)} \sim \pi$ $(i = 1, \ldots, N)$ are available to us then introduce the following *empirical measure*

$$\widehat{\pi}_N(x) := \frac{1}{N}\sum_{i=1}^{N} \delta_{X^{(i)}}(x).$$

- **The concentration of points in a given region of the space represents π.**
- This approach is in contrast with what is usually done in parametric statistics, *i.e.* start with samples and then introduce a distribution with an algebraic representation for the underlying population.


Values of the weights
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
-3
-2
-1
0
1
2
3
4
Values of x

Expectation of a general function

- The MC estimate of $\mathbb{E}_\pi(f(X))$ is

$$S_N(f) = \mathbb{E}_{\widehat{\pi}_N}(f(X)) = \frac{1}{N}\sum_{i=1}^{N} f(X^{(i)}).$$

- From the law of large numbers, we have $S_N(f)$

$$\lim_{N\rightarrow+\infty} S_N(f) = \mathbb{E}_\pi(f(X)) \; a.s.$$

- A good measure of the approximation quality is the variance of $S_N(f)$,

$$var\left[S_N(f)\right] = \frac{var_\pi\left[f(X)\right]}{N}$$

  and the CLT tells us that

$$\sqrt{N}\left(S_N(f) - \mathbb{E}_\pi(f(X)\right) \overset{N\rightarrow+\infty}{\rightarrow}_d \mathcal{N}\left(0, var_\pi\left[f(X)\right]\right)$$

- We have $var_\pi(f) = \mathbb{E}_\pi(f^2) - \mathbb{E}_\pi^2(f) \simeq \frac{1}{N}\sum_{i=1}^N f^2(X^{(i)}) - \left(\frac{1}{N}\sum_{i=1}^N f(X^{(i)})\right)^2$.

Marginalization and Optimization

- Similarly, if we have

$$\widehat{\pi}_N(x_1, x_2) = \frac{1}{N}\sum_{i=1}^{N}\delta_{X_1^{(i)},X_2^{(i)}}(x_1, x_2)$$

  so the marginal distribution is simply given by

$$\widehat{\pi}_N(x_1) = \frac{1}{N}\sum_{i=1}^{N}\delta_{X_1^{(i)}}(x_1)$$

- If we want to estimate $\arg\max\ \pi(x)$ and $\pi(x)$ is known up to a normalizing constant then

$$\underset{\{X^{(i)}\}}{\arg\max}\ \pi\left(x^{(i)}\right)$$

  is a reasonable although not very clever estimate.

Summary

- If you could sample easily from an arbitrary probability distribution,

then you could easily estimate all the quantities you are interested in.

- **Problem**: How do you sample from an arbitrary probability distribution???

- Known general techniques to sample from a general distribution π rely on the use of samples from an instrumental distribution (or proposal distribution) q.
  - For example : if $X \sim q$ then with f a diffeomorphism, $[Y = f(X)] \sim g(y) = q(f^{-1}(y)) \times |Jacobian(f^{-1})(y)|$.
  - However finding f such that $\pi(y) = g(y)$ requires *strong analytical tractability*; see the wonderful book by Devroye (1986).
- The most universal methods only require one to evaluate the densities π and q pointwise up to some normalizing constants.
- Rejection sampling is the standard approach.

Accept-Reject Method

- The rejection method allows one to sample according to a distribution π defined on X only known up to a proportionality constant, say $\pi(x) \propto \gamma(x)$.
- It relies on samples generated from a *proposal* distribution q on X. q might as well be known only up to a normalising constant, say $q(x) \propto q^*(x)$.
- We need $q^*(x)$ to 'dominate' $\gamma(x)$; i.e.

$$C = \sup_{x \in \mathcal{X}} \frac{\gamma(x)}{q^*(x)} < +\infty$$

- This implies $\gamma(x) > 0 \Rightarrow q^*(x) > 0$ but also that the tails of $q^*(x)$ must be thicker than the tails of $\gamma(x)$.

Consider $C' \geq C$. Then the accept/reject procedure proceeds as follows.

1. Sample $Y \sim q$ and $U \sim \mathcal{U}\left[0, C' q^* \left(Y\right)\right]$.
2. If $U < \gamma\left(Y\right)$ then return Y; otherwise return to step 1.

- We have for any $x \in \mathsf{X}$

$$\mathbb{P}\left(Y \leq x \text{ and } Y \text{ accepted}\right) = \int_0^{C'q^*(y)} \int_{-\infty}^{x} \mathbb{I}\left(u \leq \gamma\left(y\right)\right) q\left(y\right) \frac{1}{C'q^*(y)} dydu$$
$$= \int_{-\infty}^{x} \frac{\gamma(y)}{C'q^*(y)} q\left(y\right) dy = \frac{\int_{-\infty}^{x} \gamma(y) dy}{C' \int_{\mathcal{X}} q^*(y) dy}$$

- The probability of being accepted is the marginal of $\mathbb{P}\left(Y \leq x \text{ and } Y \text{ accepted}\right)$

$$\mathbb{P}\left(Y \text{ accepted}\right) = \frac{\int_{\mathcal{X}} \gamma\left(y\right) dy}{C' \int_{\mathcal{X}} q^*\left(y\right) dy}.$$

- Thus

$$\Pr\left(Y \leq x \middle| Y \text{ accepted}\right) = \frac{\Pr(Y \leq x \text{ and } Y \text{ accepted})}{\Pr(Y \text{ accepted})}$$
$$= \frac{\int_{-\infty}^{x} \gamma(y) dy}{\int_{\mathcal{X}} \gamma(y) dy} = \int_{-\infty}^{x} \pi\left(y\right) dy.$$

- The acceptance probability $\mathbb{P}\left(Y \text{ accepted}\right)$ is a measure of efficiency.
- The number of trials before accepting a candidate follows a geometric distribution

$$\begin{aligned} \mathbb{P}\left(k^{\text{th}} \text{ proposal accepted}\right) &= (1-\rho)^{k-1}\rho \\ \text{where } \rho &= \left(\frac{\int_{\mathcal{X}} \gamma\left(y\right) dy}{C' \int_{\mathcal{X}} q^{*}\left(y\right) dy}\right) \end{aligned}$$

thus its expected value is

$$\sum_{k=0}^{\infty} k\left(1-\rho\right)^{k-1}\rho = \frac{1}{\varrho} = \frac{1}{\Pr\left(Y \text{ accepted}\right)}.$$

- This is important to better understand the Metropolis-Hastings algorithm.

Limitations of the approach

- Consider the case where X$= \mathbb{R}^{n_x}$

$$\pi(x) = \frac{1}{(2\pi)^{n_x/2}} \exp\left(-\frac{\sum_{i=1}^{n} x_i^2}{2}\right)$$

and

$$q(x) = \frac{1}{(2\pi\sigma^2)^{n_x/2}} \exp\left(-\frac{\sum_{i=1}^{n} x_i^2}{2\sigma^2}\right) \quad .$$

- We have for any $\sigma^2 > 1$

$$\frac{\pi(x)}{q(x)} = \sigma^{n_x} \exp\left(-\frac{1}{2}\left(1-\frac{1}{\sigma^2}\right)\sum_{i=1}^{n_x} x_i^2\right) \leq \sigma^{n_x} \text{ for any } x \in \mathbb{R}^{n_x}$$

(which is reached for $x = 0$).

- Consequently

$$\mathbb{P}(Y \text{ accepted}) = \frac{1}{\sigma^{n_x}}$$

i.e. the acceptance probability decreases exponentially fast with n_x.

- Rejection Sampling is limited to problems of moderate dimensions.
- **Problem**: We try to sample all the components of a potentially high-dimensional parameter simultaneously/sequentially and we can never correct for components already sampled.
- An idea might consists of iteratively sampling components of large vectors.
- A powerful class of valid methods is available to deal with such methods: Markov chain Monte Carlo.

The nuclear pump data

- Multiple failures in a nuclear plant

Pump k	1	2	3	4	5
# Failures p_k	5	1	5	14	3
Times t_k	94.32	15.72	62.88	125.76	5.24

Pump k	6	7	8	9	10
# Failures p_k	19	1	1	4	22
Times t_k	31.44	1.05	1.05	2.10	10.48

- Model: # of failures of the $k-$th pump follow a Poisson process with parameter λ_k $(1 \leq k \leq 10)$. For an observed time t_k, the number of failures p_k is thus a Poisson $\mathcal{P}(\lambda_k t_k)$ random variable.
- The unknown parameters consist of $x = (\lambda_1, \ldots, \lambda_{10}, \beta)$ with β a parameter of the law of the $\{\lambda_k\}$.

Posterior distribution

- Hierarchical model

$$\lambda_k |\, (\alpha, \beta) \overset{\text{iid}}{\sim} \mathcal{G}a(\alpha, \beta) \text{ and } \beta \sim \mathcal{G}a(\gamma, \delta)$$

with $\alpha = 1.8$ and $\gamma = 0.01$ and $\delta = 1$.

- With $p := (p_1, \ldots p_{10})$ and $t := (t_1, \ldots t_{10})$ the posterior distribution is proportional to

$$\begin{aligned} & p\left(\lambda_1, \ldots, \lambda_{10}, \beta | p, t\right) \\ & \propto \prod_{k=1}^{10} \{\lambda_k t_k)^{p_k} \exp(-\lambda_k t_k) \lambda_k^{\alpha-1} \exp(-\beta\lambda_k\} \beta^{10\alpha} \beta^{\gamma-1} \exp(-\delta\beta) \\ & \qquad \propto \prod_{k=1}^{10} \{\lambda_k^{p_k+\alpha-1} \exp(-(t_k+\beta)\lambda_k)\} \beta^{10\alpha+\gamma-1} \exp(-\delta\beta) \ . \end{aligned}$$

- This multidimensional distribution is rather complex. It is not obvious how the rejection method or importance sampling could be used in this context.

However

- The conditionals have a familiar form

$$p(\lambda_1, \ldots, \lambda_{10} | p, t, \beta) = \prod_{k=1}^{10} p(\lambda_k | p_k, t_k, \beta)$$

  where

$$\lambda_k | p_k, t_k, \beta \sim \mathcal{G}a(p_k + \alpha, t_k + \beta) \text{ for } 1 \leq k \leq 10,$$

  and

$$\beta | (\lambda_1, \ldots, \lambda_{10}) \sim \mathcal{G}a(\gamma + 10\alpha, \delta + \sum_{k=1}^{10} \lambda_k) .$$

- Instead of directly sampling the vector $x = (\lambda_1, \ldots, \lambda_{10}, \beta)$ at once, one could suggest sampling it iteratively, starting for example with the λ_i's for a given guess of β, followed by an update of β given the new samples $\lambda_1, \ldots, \lambda_{10}$.

My first Gibbs sampler

- Given a sample, at iteration i, $x^{(i)} := (\lambda_1^{(i)}, \ldots, \lambda_{10}^{(i)}, \beta^{(i)})$ one could proceed as follows at iteration $i+1$,

1. $\lambda_k^{(i+1)} | (\beta^{(i)}, t_k, p_k) \sim \mathcal{G}a(p_k + \alpha, t_k + \beta^{(i)})$ for $1 \leq k \leq 10$,
2. $\beta^{(i+1)} | (\lambda_1^{(i+1)}, \ldots, \lambda_{10}^{(i+1)}) \sim \mathcal{G}a(\gamma + 10\alpha, \delta + \sum_{k=1}^{10} \lambda_k^{(i+1)})$.

- Instead of directly sampling in a space with 11 dimensions, one samples in spaces of dimension 1.
- One could as well choose randomly at each iteration the parameter among $\lambda_1, \ldots, \lambda_{10}, \beta$ to update.

Many questions

- The structure of the algorithm calls for many questions:
  - Are we sampling from the desired joint distribution?
  - If yes, how many times should the iteration above be repeated?
- The validity of the approach described here stems from the fact that the sequence $\{X^{(i)}\}$ defined above is a Markov chain and some Markov chains have very nice properties.

Elements of Markov chains

- **Markov chain**: A sequence of random variables $\left\{X^{(n)}; n \in \mathbb{N}\right\}$ defined on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ which satisfies the property, for any $A \in \mathcal{B}(\mathsf{X})$

$$\mathbb{P}\left(X^{(n)} \in A \middle| X_0, ..., X^{(n-1)}\right) = \mathbb{P}\left(X^{(n)} \in A \middle| X^{(n-1)}\right).$$

  and we will denote

$$P(x, A) = \int_A P(x, dy) := \mathbb{P}\left(X^{(n)} \in A \middle| X^{(n-1)}\right) \ .$$

- Note that the marginal joint distribution of $(X^{(n-1)}, X^{(n)})$ for $n \geq 1$ is

$$\begin{aligned}
&\mathbb{P}\left(X^{(n-1)} \in A, X^{(n)} \in B\right) = \int_A \mathbb{P}\left(X^{(n-1)} \in dx, X^{(n)} \in B\right) \\
&= \int_A \mathbb{P}\left(X^{(n-1)} \in dx\right) \mathbb{P}\left(X^{(n)} \in B | X^{(n-1)} = x\right) \\
&= \int_A \mathbb{P}\left(X^{(n-1)} \in dx\right) P(x, B) \ .
\end{aligned}$$

- **Markov chain Monte Carlo**: Given a target π, design a transition kernel P such that asymptotically as $n \to \infty$

$$\frac{1}{N}\sum_{n=1}^{N}\varphi\left(X^{(n)}\right) \rightarrow \int_{\mathsf{X}}\varphi\left(x\right)\pi\left(x\right)dx \text{ and/or } X^{(n)} \sim \pi.$$

- It should be easy to simulate the Markov chain even if π is complex.
- Intuitively we should require that if $X^{(n-1)} \sim \pi$ then also $X^{(n)} \sim \pi$, that is mathematically

$$\begin{aligned}
\mathbb{P}\left(X^{(n)} \in B\right) &= \pi(X^{(n)} \in B) \\
&= \mathbb{P}\left(X^{(n-1)} \in \mathsf{X}, X^{(n)} \in B\right) \\
&= \int_{\mathsf{X}} \mathbb{P}\left(X^{(n-1)} \in dx\right) P\left(x, B\right) \\
&= \int_{\mathsf{X}} \pi(dx) P\left(x, B\right) \ .
\end{aligned}$$

Normal autoregressive example

- Consider the autoregression for $|\alpha| < 1$

$$X^{(n)} = \alpha X^{(n-1)} + V^{(n)}, \text{ where } V^{(n)} \sim \mathcal{N}\left(0, \sigma^2\right)$$

then

$$P\left(x, dy\right) = P\left(x, y\right) dy = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{\left(y - \alpha x\right)^2}{2\sigma^2}\right) dx.$$

- One can easily check that

$$\int_{\mathsf{X}} \pi\left(x\right) P\left(x, y\right) dx = \pi\left(y\right)$$

with

$$\pi\left(x\right) = \mathcal{N}\left(x; 0, \frac{\sigma^2}{1 - \alpha^2}\right) \ .$$

- To sample from π, we could just sample the Markov chain and asymptotically we would have $X^{(n)} \sim \pi$. [Obviously, in this case this is useless because we can sample from π directly.]

Illustration

- Graphically, consider 1000 independent Markov chains run in parallel.
- We assume that the initial distribution of these Markov chains is $\mathcal{U}_{[0,20]}$. So initially, the Markov chains samples are not distributed according to π.

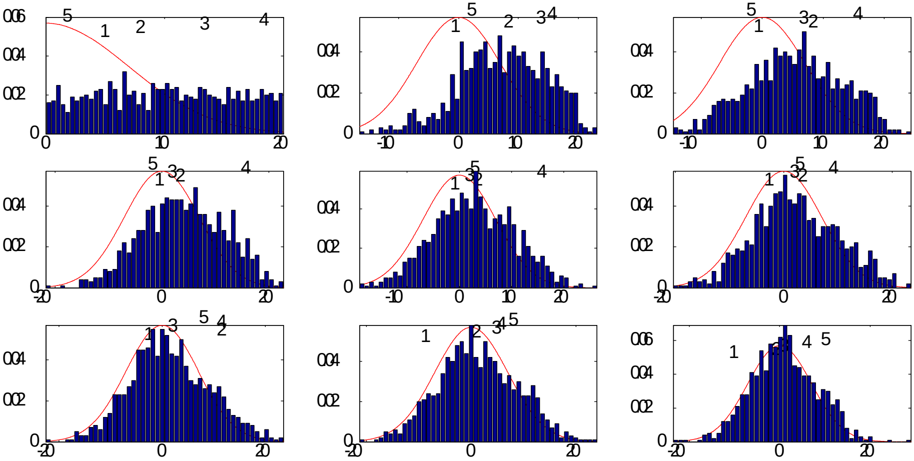


Figure: From top left to bottom right: histograms of 1000 independent Markov chains with a normal distribution as target distribution as *n* increases.

- The target normal distribution seems to "attract" the distribution of the samples and even to be a fixed point of the algorithm.
- Once close to π the histogram never drifts away.
- This is is what we wanted to achieve, *i.e.* it seems that we have produced 1000 independent samples from the normal distribution.
- In fact one can show that it is not necessary to run N Markov chains in parallel in order to obtain 1000 samples, but that one can consider a unique Markov chain, and build the histogram from this single Markov chain by forming histograms from one trajectory.

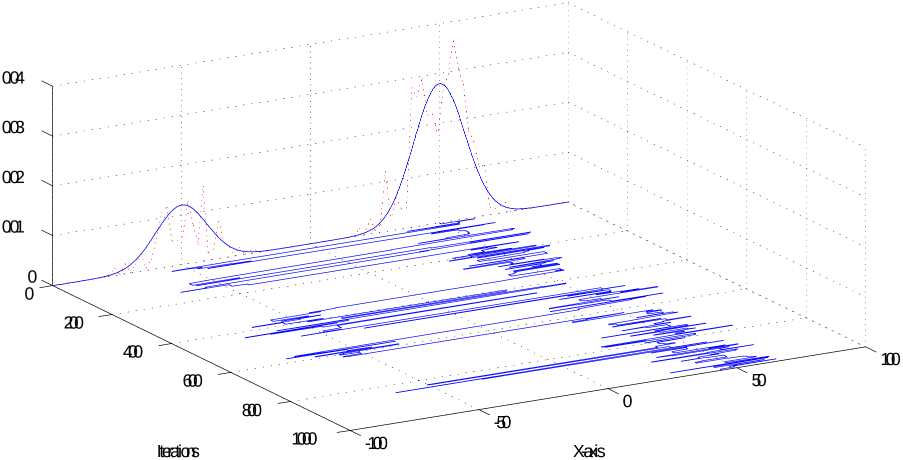


Figure: Bimodal target distributions and simulated Markov chain

Estimate of expectations

- The estimate of the target distribution, through the series of histograms, improves with the number of iterations.
- Assume that we have stored $\{X^{(n)}, 1 \leq n \leq N\}$ for N large and wish to estimate $\int_{\mathbb{X}} \varphi(x)\pi(x)dx$.
- In the light of the numerical experiments, one can suggest the estimator

  $$\frac{1}{N}\sum_{n=1}^{N} \varphi(X^{(n)}) \ .$$

  which is exactly the estimator that we would use if $\{X^{(n)}, 1 \leq n \leq N\}$ were independent.
- In fact, it can be proved, under relatively mild conditions, that such an estimator is consistent *despite the fact that the samples are NOT independent.* Under additional conditions, a CLT also holds with a rate of CV in $1/\sqrt{N}$.

Summary

To summarize, we are interested in Markov chains with transition kernel P which have the following three important properties observed above:

- The desired distribution π is a "fixed point" of the algorithm or, in more appropriate terms, an *invariant distribution* of the Markov chain, *i.e.* $\int_{\mathsf{X}} \pi(x)P(x,y)dx = \pi(y)$.
- The estimator $\frac{1}{N}\sum_{n=1}^{N}\varphi(X^{(n)})$ converges towards $\mathbb{E}_{\pi}(\varphi(X))$
- The successive distributions of the Markov chains converge towards π; i.e. asymptotically $X^{(n)} \sim \pi$

As we shall see

- Given $\pi(x)$, there is an infinite number of kernels $P(x, y)$ which have $\pi(x)$ as their invariant distribution.
- Convergence is ensured under very weak assumptions; namely irreducibility and aperiodicity.
- The "art" of MCMC consists of constructing "efficient" transitions.
- However it is usually very easy to establish that an MCMC sampler converges towards π but very difficult to obtain rates of convergence.

Two fundamental properties

- Let P_1 and P_2 be two Markov transition probabilities with common invariant distribution π *i.e.* for $i = 1, 2$

$$\int_{\mathsf{X}} \pi(x) P_i(x, y) dx = \pi(y),$$

then,

1. if we assume $X^{(n-1)} \sim \pi$ and let $\tilde{X}^{(n)} \sim P_1(X^{(n-1)}, \cdot)$ and $X^{(n)} \sim P_2(\tilde{X}^{(n)}, \cdot)$, then $X^{(n)} \sim \pi$, that is the *composition* $P := P_1 P_2$ leaves π invariant, since

$$\int_{\mathsf{X}} \left[\int_{\mathsf{X}} \pi(x) P_1(x, y) dx \right] P_2(y, z) dy = \int_{\mathsf{X}} \pi(y) P_2(y, z) dy = \pi(z)$$

2. if we assume $X^{(n-1)} \sim \pi$ and choose P_1 with probability $\lambda \in [0, 1]$ (resp. P_2 with probability $1 - \lambda$) and let $X^{(n)} \sim P_1(X^{(n-1)}, \cdot)$ (resp. $X^{(n)} \sim P_1(X^{(n-1)}, \cdot)$) then $X^{(n)} \sim \pi$, that is the *mixture* $P := \lambda P_1 + (1 - \lambda) P_2$ leaves π invariant, since

$$\begin{aligned} &\textstyle\int_{\mathsf{X}} \pi(x) \left[\lambda P_1(x, y) + (1 - \lambda) P_2(x, y) \right] dx \\ &= \lambda \textstyle\int_{\mathsf{X}} \pi(x) P_1(x, y) + (1 - \lambda) \int_{\mathsf{X}} \pi(x) P_2(x, y) dx = \pi(y) \end{aligned}$$

Application to the two component Gibbs sampler

- Consider the target distribution $\pi(x)$ such that $x = (x_1, x_2) \in \mathsf{X}^2$. Then the 2-component Gibbs sampler proceeds as follows.
- Initialization: Select deterministically or randomly $x^{(0)} = \left(x_1^{(0)}, x_2^{(0)}\right)$.
- Iteration i; $i \geq 1$
  - Sample $X_1^{(i)} \sim \pi\left(\cdot | \, x_2^{(i-1)}\right)$.
  - Sample $X_2^{(i)} \sim \pi\left(\cdot | \, x_1^{(i)}\right)$.
- Sampling from these conditional is often feasible even when sampling from the joint is impossible (*e.g.* nuclear pump data).

Invariance

- The algorithm is the composition of two Markov transition probabilities

$$\begin{aligned}P_1(x_1, x_2; y_1, y_2) &= \pi(y_1|x_2)\delta_{x_2}(y_2)\\ P_2(x_1, x_2; y_1, y_2) &= \pi(y_2|x_1)\delta_{x_1}(y_1)\end{aligned}$$

- Then we have

$$\begin{aligned}&\int_X \int_X \pi(x_1, x_2) P_1(x_1, x_2; y_1, y_2)\, dx_1\, dx_2\\ &\qquad = \int_X \int_X \pi(x_1, x_2)\pi(y_1|x_2)\delta_{x_2}(y_2)\, dx_1\\ &\qquad\qquad = \int_X \pi(x_2)\pi(y_1|x_2)\delta_{x_2}(y_2)\, dx_2 = \pi(y_1, y_2)\ .\end{aligned}$$

- Conclusion the Gibbs sampler leaves π invariant!

Irreducibility

- This does not ensure that the Gibbs sampler does converge towards the invariant distribution!
- Additionally it is required to ensure *irreducibility*: loosely speaking the Markov chain can move to any set A such that $\pi(A) > 0$ for (almost) any starting point.
- This ensures that

$$\frac{1}{N}\sum_{n=1}^{N}\varphi\left(X_1^{(n)}, X_2^{(n)}\right) \rightarrow \int \varphi(x_1, x_2)\,\pi(x_1, x_2)\,dx_1\,dx_2$$

but NOT that asymptotically $\left(X_1^{(n)}, X_2^{(n)}\right) \sim \pi$.

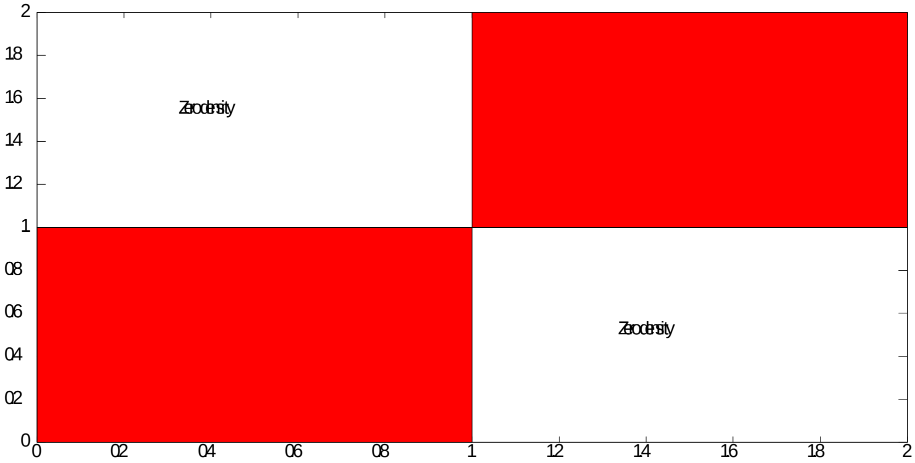


Figure: A distribution that can lead to a reducible Gibbs sampler.

Aperiodicity

- Consider a simple example where $\mathsf{X} = \{1, 2\}$ and $P(1, 2) = P(2, 1) = 1$. Clearly the invariant distribution is given by $\pi(1) = \pi(2) = \frac{1}{2}$.
- However, we know that if the chain starts in $X^{(0)} = 1$, then $X^{(2n)} = 1$ and $X^{(2n+1)} = 0$ for any n.
- We have

$$\frac{1}{N}\sum_{n=1}^{N} \varphi\left(X^{(n)}\right) \rightarrow \int \varphi(x)\, \pi(x)\, dx$$

  but clearly $X^{(n)}$ is NOT distributed according to π.
- One needs to make sure that you do NOT explore the space in a periodic way to ensure that $X^{(n)} \sim \pi$ asymptotically.

Deterministic Scan Gibbs Sampler

- If $x = (x_1, ..., x_p)$ where $p \geq 2$, the Gibbs sampling strategy still applies.
- Initialization: Select deterministically or randomly $x^{(0)} = \left(x_1^{(0)}, ..., x_p^{(0)}\right)$.
- Iteration i; $i \geq 1$:
  - For $k = 1 : p$
    - Sample $X_k^{(i)} \sim \pi\left(\cdot \,|\, x_{-k}^{(i)}\right)$ where $x_i^{-k} = \left(x_1^{(i)}, ..., x_{k-1}^{(i)}, x_{k+1}^{(i-1)}, ..., x_p^{(i-1)}\right)$.

Random Scan Gibbs Sampler

- Initialization: Select deterministically or randomly $x^{(0)} = \left(x_1^{(0)}, ..., x_p^{(0)} \right)$.
- Iteration i; $i \geq 1$:
    - Sample $K \sim U_{\{1,...,p\}}$.
    - Sample $X_K^{(i)} \sim \pi \left(\cdot | \, x_{-K}^{(i)} \right)$ where $x_{-K}^{(i)} = \left(x_1^{(i-1)}, ..., x_{K-1}^{(i-1)}, x_{K+1}^{(i-1)}, ..., x_p^{(i-1)} \right)$.

Random Scan Gibbs Sampler

- Initialization: Select deterministically or randomly $x^{(0)} = \left(x_1^{(0)}, ..., x_p^{(0)} \right)$.
- Iteration i; $i \geq 1$:
  - Sample $K \sim U_{\{1,...,p\}}$.
  - Sample $X_K^{(i)} \sim \pi \left(\cdot \middle| \, x_{-K}^{(i)} \right)$ where $x_{-K}^{(i)} = \left(x_1^{(i-1)}, ..., x_{K-1}^{(i-1)}, x_{K+1}^{(i-1)}, ..., x_p^{(i-1)} \right)$.

Nuclear pumps - again

- At iteration $i+1$,

1. $\lambda_k^{(i+1)} | (\beta^{(i)}, t_k, p_k) \sim \mathcal{G}a(p_k + \alpha, t_k + \beta^{(i)})$ for $1 \leq k \leq 10$,
2. $\beta^{(i+1)} | (\lambda_1^{(i+1)}, \ldots, \lambda_{10}^{(i+1)}) \sim \mathcal{G}a(\gamma + 10\alpha, \delta + \sum_{k=1}^{10} \lambda_k^{(i+1)})$.

- The conditionals have positive density on $[0, +\infty)$: this implies *irreducibility*.
- For the same reason no periodic behaviour is possible : *aperiodicity*.
- Conclusion: it is a theoretically valid algorithm!
- However conditioning comes at a price.

Yet another toy example

- Target

$$\pi(x,y) = \mathcal{N}\left((x,y)^{\mathrm{T}}; 0, \Sigma\right) \text{ with } \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

  where $\rho \in (-1,1)$ is the correlation coefficient between x and y.

- The conditional distributions $\pi(x|y)$ and $\pi(y|x)$ are therefore

$$\pi(x|y) = \mathcal{N}\left(x; \rho y, (1-\rho^2)\right)$$

  and

$$\pi(y|x) = \mathcal{N}\left(y; \rho x, (1-\rho^2)\right).$$

- Hence with $V_1^{(n)}, V_2^{(n)} \sim \mathcal{N}(0, I_2)$,

$$\begin{aligned} X^{(n+1)} &= \rho Y^{(n)} + \sqrt{1-\rho^2} V_1^{(n+1)} \\ Y^{(n+1)} &= \rho X^{(n+1)} + \sqrt{1-\rho^2} V_2^{(n+1)}. \end{aligned}$$

- Observe that when $\rho \to 1$ then $X^{(n+1)} \simeq Y^{(n)}$ and $Y^{(n+1)} \simeq X^{(n+1)}$.

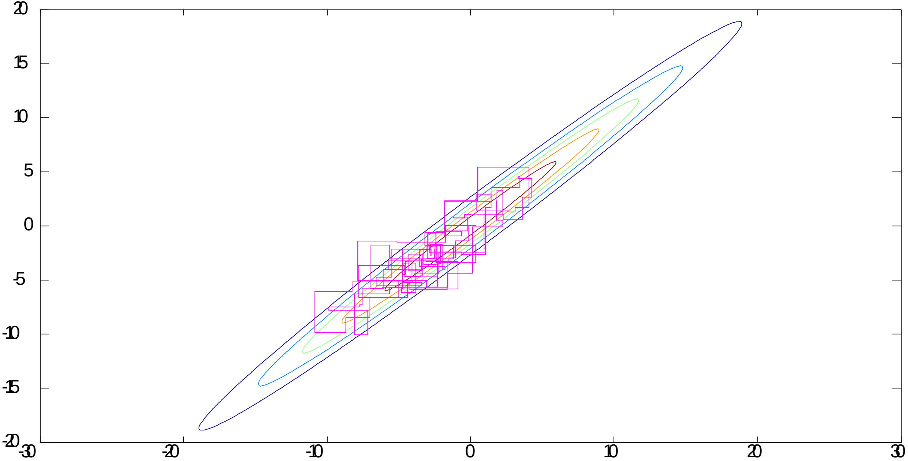


Figure: Even when irreducibility and aperiodicity are ensured, the Gibbs sampler can still converge very slowly.

Discussion

- The Gibbs sampler is a rather generic tool to sample approximately from high-dimensional distributions.
- It allows one to break a large and difficult sampling problem into smaller, often more tractable, sampling sub-problems.
- However the approach raises new challenges: there exists a tension between
  1. the potentially detrimental effect of not updating dependent components simultaneously,
  2. the fact that it is usually easier to sample from lower dimensional distribution (*e.g.* assume that a rejection algorithm is used to sample from the $\mathcal{G}a$ distribution involved in the Gibbs sampler for the nuclear pump data)
- This is a difficulty for most Monte Carlo methods.

Tricks of the trade

- Try to have as few “blocks” as possible.
- Put the most correlated variables in the same block.
- If necessary, reparametrise the model to achieve this.
- Integrate analytically as many variables as possible.
- However with the Gibbs sampler one is heavily constrained by the structure of the target and tractability issues,
    1. the only degree of freedom is the choice of the partition,
    2. the Metropolis-Hastings algorithm allows one to circumvent this lack of flexibility.

Difficulties with the Gibbs sampler

- The Gibbs sampler requires sampling from the full conditional distributions
$$\pi\left(x_k \middle| x_{-k}\right).$$
- For many complex models, it is impossible to sample from several of these "full" conditional distributions.
- Even if it is possible to implement the Gibbs sampler, the algorithm might be very inefficient because the variables are very correlated or sampling from the full conditionals is extremely expensive/inefficient.

Metropolis-Hastings Algorithm

- The Metropolis-Hastings algorithm is a general strategy to construct Markov transition probabilities with a given invariant distribution $\pi(x)$.
- It is in fact the main building block of MCMC algorithms and provides extreme flexibility.
- The Metropolis algorithm was named the "Top algorithm of the 20th century" by computer scientists, mathematicians and physicists.

Ingredients

- Introduce a family of proposal distribution $\{q(x,\cdot), x \in \mathsf{X}\}$, *i.e.*

$$\int q(x,y)\, dy = 1 \text{ for any } x \in \mathsf{X} .$$

- The basic idea of the MH algorithm is,
  1. given that the current state of the Markov chain is $x \in \mathsf{X}$, to propose a new candidate y from $q(x,\cdot)$,
  2. to accept the proposed sampled with an appropriate probability $\alpha(x,y)$ which ensures that the invariant distribution of the transition kernel is the target distribution $\pi(x)$.

The Metropolis-Hastings update

- Initialization: Select deterministically or randomly x_0.
- Iteration i; $i \geq 1$:
    - Sample $y \sim q\left(x^{(i-1)}, \cdot\right)$ and compute

$$\alpha\left(x^{(i-1)}, y\right) = \min\left\{1, \frac{\pi(y)\, q\left(y, x^{(i-1)}\right)}{\pi\left(x^{(i-1)}\right) q\left(x^{(i-1)}, y\right)}\right\}.$$

    - With probability $\alpha\left(x^{(i-1)}, y\right)$, set $x^{(i)} = y$; otherwise set $x^{(i)} = x^{(i-1)}$.

Reversibility aka Detailed Balance

- One can easily show that the M-H kernel is π-reversible

  $$\pi(x) K(x,y) = \pi(y) K(y,x)$$

  where

  $$K(x,y) = q(x,y)\,\alpha(x,y) + \left(1 - \int q(x,y')\,\alpha(x,y')\,dy'\right)\delta_x(y).$$

- Indeed we have

  $$\begin{aligned} \pi(x)\,q(x,y)\,\alpha(x,y) &= \pi(x)\,q(x,y)\min\left\{1, \frac{\pi(y)\,q(y,x)}{\pi(x)\,q(x,y)}\right\} \\ &= \min\{\pi(x)\,q(x,y), \pi(y)\,q(y,x)\} \\ &= \pi(y)\,q(y,x)\,\alpha(y,x) \end{aligned}$$

- π-reversibily implies straightforwardly π-invariance.

Irreducibility and Aperiodicity

- To ensure irreducibility, a sufficient but not necessary condition is that

$$\pi\left(y\right) > 0 \Rightarrow q\left(x, y\right) > 0.$$

- Aperiodicity is automatically ensured as there is always a strictly positive probability to reject the candidate.
- Theoretically, the MH algorithm converges under very weak assumptions to the target distribution π. In practice, this convergence can be so slow that the algorithm is useless.

Remarks

- It is only necessary to know $\pi(x)$ up to a normalizing constant to implement the algorithm.
- This algorithm is extremely general: $q(x, \cdot)$ can be any proposal distribution. So in practice, we can select it so that it is easy to sample from.
- There is (potentially) much more freedom than with the Gibbs sampler where
  1. the proposal distributions are constrained by the structure of π
  2. the tractability of some of the conditional distributions.

Independent Metropolis-Hastings

- Consider the simple choice

  $$q(x,y) = q(y);$$

  i.e. this is a so-called *independent proposal*.

- In this case, the acceptance probability is given by

  $$\alpha(x,y) = \min\left\{1, \frac{\pi(y)\,q(x)}{\pi(x)\,q(y)}\right\} = \min\left\{1, \frac{\gamma(y)}{q^{*}(y)}\frac{q^{*}(x)}{\gamma(x)}\right\}$$

  where γ and q^{*} are unnormalised versions of π and q.

- The ratio $\gamma(x)/q^{*}(x)$ appearing in the Accept/Reject also reappears here.
- Given this resemblance, one might wonder if this is likely to be a good approach?

Example

- **Example**: Consider the case where

$$\pi(x) \propto \exp\left(-\frac{x^2}{2}\right).$$

- We implement the MH algorithm for

$$q_1(x) \propto \exp\left(-\frac{x^2}{2(0.2)^2}\right)$$

so $\pi(x)/q_1(x) \to \infty$ as $x \to \infty$ and for

$$q_2(x) \propto \exp\left(-\frac{x^2}{2(5)^2}\right)$$

so $\pi(x)/q_2(x) \leq C < \infty$ for all x.

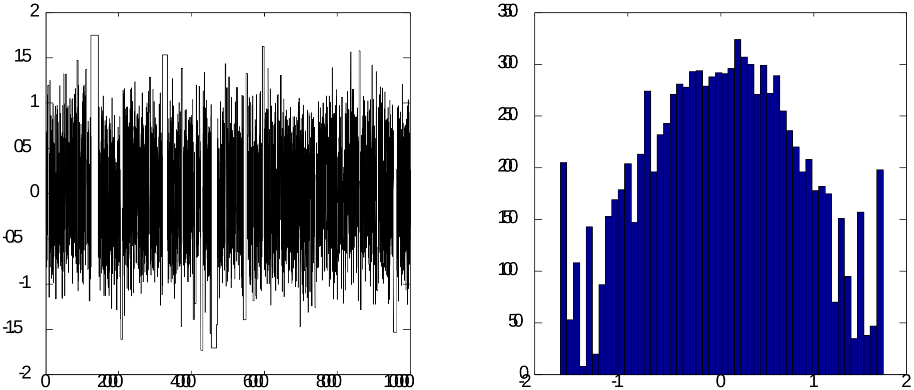


Figure: MCMC output for q_1, we estimate $\mathbb{E}\left(X\right) = 0.0206$ and $\mathbb{V}\left(X\right) = 0.83$.

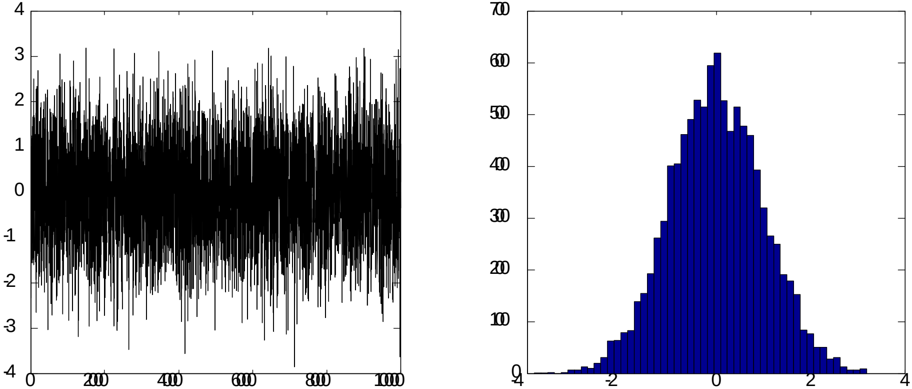


Figure: MCMC output for q_2, we estimate $\mathbb{E}\left(X\right) = -0.004$ and $\mathbb{V}\left(X\right) = 1.00$.

Problem exacerbated by high-dimension

- Again consider the example from the IS introduction:
  - $\pi(x) = \mathcal{N}(x; 0, I)$ with $x \in \mathbb{R}^{n_x}$.
  - $q(x) = \mathcal{N}(x; \varepsilon \times e, I)$ with $e = (1, 1, 1 \ldots)^{\mathrm{T}}$.
- We have

$$\begin{aligned}\frac{\pi(x)}{q(x)} &= \exp\left(\frac{1}{2} n_x - \varepsilon e^{\mathrm{T}} x\right) \\ &= \exp\left(\frac{1}{2}\varepsilon^2 n_x - \varepsilon\sqrt{n_x}\frac{1}{\sqrt{n_x}} \sum_{i=1}^{n_x} x(i)\right)\end{aligned}$$

  which suggests a high variability of the weights as n_x increases.
- For $\varepsilon = 1$, $n_x = 10$ and $N = 10,000$ we observe and acceptance rate of $\simeq 0.3\%$ and $\left|\frac{1}{N n_x} \sum_{i=1}^{N} \sum_{k=1}^{n_x} x_k^{(i)}\right| = 5.43$.

Divide and conquer...

- This time,
  - $\pi(x) = \mathcal{N}(x; 0, I)$ with $x \in \mathbb{R}^{n_x}$.
  - $q(x) = \mathcal{N}(x; \varepsilon \times e, I)$ with $e = (1, 1, 1 \ldots)^{\mathrm{T}}$ and Σ a $n_x \times n_x$ covariance matrix.
- One can suggest the following algorithm, at iteration $i + 1$,
  1. Choose a coordinate $k \sim \mathcal{U}\{1, 2, \ldots, n_x\}$
  2. Propose $y_k \sim \mathcal{N}(\varepsilon, 1) =: q_k$ (the marginal of q above).
  3. Set $x_{-k}^{(i+1)} = x_{-k}^{(i)}$ and $x_k^{(i+1)} = y_k$ with probability

$$\min\left\{1, \frac{\pi(y)q_k(x_k^{(i)})}{\pi(x^{(i)})q_k(y_k)}\right\} = \min\left\{1, \frac{\pi(y_k|x_{-k}^{(i)})\pi(x_{-k}^{(i)})q_k(x_k^{(i)})}{\pi(x_k^{(i)}|x_{-k}^{(i)})\pi(x_{-k}^{(i)})q_k(y_k)}\right\}$$

- For $\varepsilon = 1$, $n_x = 10$ and $N = 10,000$ we obtain an acceptance rate of $\simeq 50\%$ and $\left|\frac{1}{Nn_x}\sum_{i=1}^{N}\sum_{k=1}^{n_x} x_k^{(i)}\right| \simeq 1.60$.

- However with $\Sigma \neq I$ ($\Sigma = CC^{\mathrm{T}}$ with $C_{ij} \sim \mathcal{N}(0,1)$ and *iid*)
    - Full update: acceptance rate $\simeq 0.6\%$ and $\left|\frac{1}{Nn_x}\sum_{i=1}^{N}\sum_{k=1}^{n_x} x_k^{(i)}\right| \simeq 28.0$
    - One at time update: acceptance rate $\simeq 40\%$ $\left|\frac{1}{Nn_x}\sum_{i=1}^{N}\sum_{k=1}^{n_x} x_k^{(i)}\right| \simeq 29.0$.

Discussion

- When using independent proposals then the right criterion is $q(x) \approx \pi(x)$ - a high acceptance rate is desirable.
- As for Rejection sampling or Importance Sampling, it is a good idea to have
$$\frac{\pi(x)}{q(x)} \leq C$$
to obtain good performance.
- These two conditions are usually difficult to ensure in practice, especially for n_x large.
- The MC malediction strikes again:
    - bold moves are desirable, but difficult to achieve
    - while timid moves are possible, but somehow inefficient.

Random Walk Metropolis

- The original Metropolis algorithm (1953) corresponds to the following choice for $q(x, y)$

  $$y = x + Z \text{ where } Z \sim f;$$

  i.e. this is a so-called *random walk proposal*.

- The distribution $f(z)$ is the distribution of the random walk increments Z and

  $$q(x, y) = f(y - x) \quad \Rightarrow \; \alpha(x, y) = \min\left\{1, \frac{\pi(y) f(x - y)}{\pi(x) f(y - x)}\right\}.$$

- If $f(y - x) = f(x - y)$ - e.g. $Z \sim \mathcal{N}(0, \Sigma)$- then

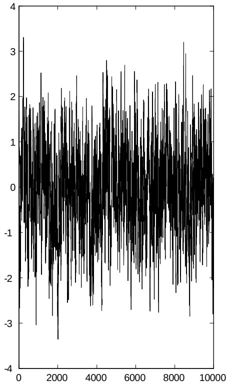
  $$\alpha(x, y) = \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}$$

- **Example**: Consider the case where

$$\pi\left(x\right)\propto\exp\left(-\frac{x^{2}}{2}\right).$$

- We implement the MH algorithm for

$$\begin{aligned}
q_1\left(x,y\right) &\propto \exp\left(-\frac{\left(y-x\right)^{2}}{2\left(0.2\right)^{2}}\right),\\
q_2\left(x,y\right) &\propto \exp\left(-\frac{\left(y-x\right)^{2}}{2\left(5\right)^{2}}\right),\\
q_3\left(x,y\right) &\propto \exp\left(-\frac{\left(y-x\right)^{2}}{2\left(0.02\right)^{2}}\right).
\end{aligned}$$



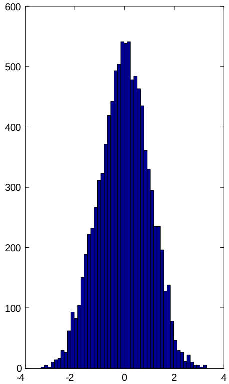
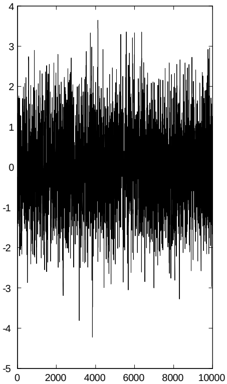


Figure: MCMC output for q_1, we estimate $\mathbb{E}(X) = -0.02$ and $\mathbb{V}(X) = 0.99$



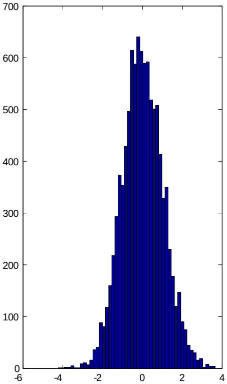


Figure: MCMC output for q_2, we estimate $\mathbb{E}(X) = 0.00$ and $\mathbb{V}(X) = 1.02$.

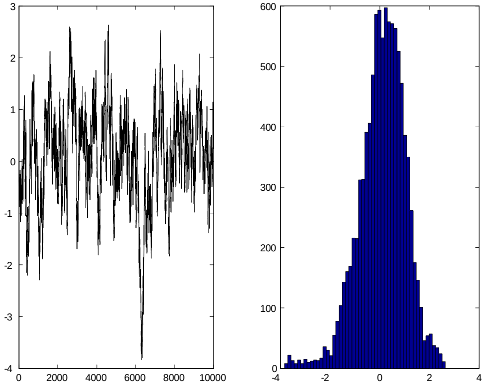


Figure: MCMC output for q_3, we estimate $\mathbb{E}(X) = 0.10$ and $\mathbb{V}(X) = 0.92$.

- A large acceptance probability is not necessarily a good criterion.
- When the variance of the random walk increments (if it exists) is very small then the acceptance rate can be expected to be around 0.5.
- One would like to scale the random walk moves such that it is possible to move reasonably fast in regions of positive probability masses under π.
- The multivariate case is even more complex since the correlation structure of the proposal and target distribution should be similar.

Limitations of the MH algorithm

- The MH algorithm is a simple and very general algorithm to sample from a target distribution $\pi(x)$.
- In practice, the choice of the proposal distribution is paramount to obtain an efficient algorithm.
- Whereas timid moves are easily designed, bold and efficient moves are very difficult to design in practice.
- A way to learn from the past is to build adaptive MCMC samplers.

Advanced MCMC Methods

- Active area for 50 years...
- Most methods relie on the introduction of auxiliary variables and associated target distributions to ease the sampling task.
- Parallel tempering, slice sampling, Hamiltonian Monte Carlo, Wang-Landau.
- Normalizing constant estimates estimation: Bridge sampling, Path sampling, AIS etc.

Slice Sampling

- Consider an initial target distribution $\pi(x) = \gamma(x) / Z$ and the extended target
$$\widetilde{\pi}(x, u) \propto \mathbb{I}_{(0,\gamma(x))}(u)$$
- We have
$$\widetilde{\pi}(x) \propto \int \mathbb{I}_{(0,\gamma(x))}(u)\, du = \gamma(x)$$
so $\widetilde{\pi}(x) = \pi(x)$.
- Moreover we have
$$\widetilde{\pi}(u|x) = \frac{\mathbb{I}_{(0,\gamma(x))}(u)}{\gamma(x)}$$
and
$$\widetilde{\pi}(x|u) \propto \begin{cases} 1 & \text{if } \gamma(x) \geq u \\ 0 & \text{otherwise.} \end{cases}$$

Hamiltonian Monte Carlo

- Assume you want to sample from $\pi(x)$ where $x \in \mathbb{R}^n$.
- Consider the extended target distribution where v is the 'velocity'

$$\widetilde{\pi}(x, v) \propto \gamma(x) \mathcal{N}(v; 0, \mathbb{I})$$

- The Hamiltonian is defined as

$$H(x, v) = \underbrace{-\log \gamma(x)}_{E(x)} + v^{\mathsf{T}} v/2.$$

- Hamiltonian dynamics are deterministic dynamics such that $(x, v) \rightarrow (x', v')$ with $H(x, v) = H(x', v')$.
- Hamiltonian dynamics are time reversible $(x', -v') \rightarrow (x, -v)$.

Hamiltonian Monte Carlo

- **Ideal MCMC**: Sample $v \sim \mathcal{N}(v; 0, \mathbb{I})$ then simulate Hamiltonian dynamics.
- Problem: Hamiltonian dynamics cannot be simulated exactly. on a computer.
- Many numerical schemes - e.g. Leap Frog - have been proposed which do not conserve Hamiltonian but are still deterministic and reversible.
- They require M-H acceptance rates; i.e. $\min\left(1, \widetilde{\pi}(x', v') / \widetilde{\pi}(x, v)\right)$.

Hamiltonian Monte Carlo

- **Ideal MCMC**: Sample $v \sim \mathcal{N}(v; 0, \mathbb{I})$ then simulate Hamiltonian dynamics.
- Problem: Hamiltonian dynamics cannot be simulated exactly. on a computer.
- Many numerical schemes - e.g. Leap Frog - have been proposed which do not conserve Hamiltonian but are still deterministic and reversible.
- They require M-H acceptance rates; i.e. $\min\left(1, \widetilde{\pi}(x', v') / \widetilde{\pi}(x, v)\right)$.
- Recent development include Riemman Manifold MCMC (Girolami et al., JRSS B, 2010); see also survey by Neal (2010).

Parallel Tempering

- Parallel tempering uses parallel chains to ease sampling (Geyer & Thompson, 1990).
- Sequence of targets $\{\pi_k(x)\}_{k=1,\dots,P}$ such that $\pi_k(x) \propto [\pi(x)]^{\phi_k}$ where $\phi_1 \geq \cdots \geq \phi_P = 1$.
- We build a Markov chain $\left\{X_1^{(n)}, \dots, X_P^{(n)}\right\}_{n\geq 1}$ of invariant distribution $\prod_{k=1}^{P} \pi_k(x_k)$ using with-in chain standard π_k-invariant MCMC kernel K_k and swap moves; e.g. swap x_k and x_l with proba.

$$\min\left\{1, \frac{\pi_k(x_l)\,\pi_l(x_k)}{\pi_k(x_k)\,\pi_l(x_l)}\right\}.$$

Bayesian Inference for Mixture Models

- Model

$$Y_i \overset{\text{i.i.d.}}{\sim} \sum_{k=1}^{4} \omega_k \mathcal{N}\left(\mu_k, \lambda_k\right).$$

- Standard conditionally conjugate priors on $\theta = (\omega_{1:4}, \mu_{1:4}, \lambda_{1:4})$, no identifiability constraint

$$\mu_k \sim \mathcal{N}(\xi, \kappa^{-1}), \lambda_k \sim \mathcal{G}a(\nu, \chi),\ \omega_{1:4} \sim \mathcal{D}(\rho).$$

- The posterior is a mixture of $4! = 24$ similar components.

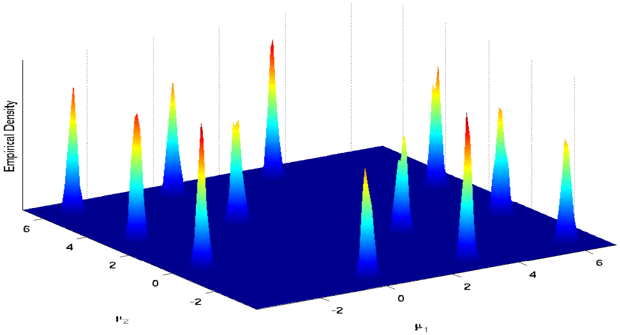
Experimental Setup

- T =100 data with $\mu = (-3, 0, 3, 6)$, $\lambda = (0.55, 0.55, 0.55, 0.55)$; components "far" from each other.
- We build the sequence of P distributions

$$\pi_n(\theta) \propto l(y_{1:T};\theta)^{\phi_n} f(\theta)$$

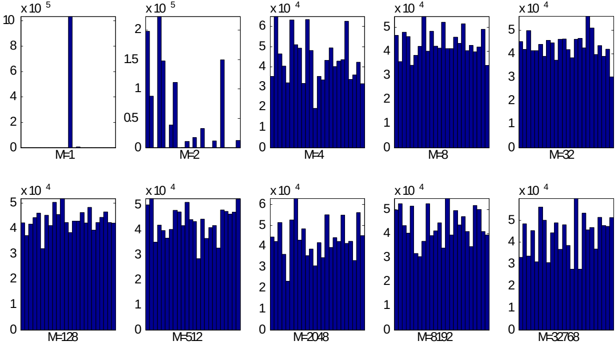
  where $\phi_1 = 0 < \phi_2 < ... < \phi_P = 1$.
- MCMC sampler to sample from π_n : update $\mu_{1:4}$ via a MH kernel with additive normal random walk, update $\lambda_{1:4}$ via a MH kernel with multiplicative log-normal random walk, update $\omega_{1:4}$ via a MH kernel with additive normal random walk on the logit scale.
- K_P admits as invariant distribution $\pi_P = \pi$. Very long runs of MCMC get trapped in one of the 4!=24 modes of the distributions.

Posterior Distribution Estimates



Marginal posterior estimated using parallel tempering with $P = 32768$ and $N = 1048576$.

Number of Samples Per Mode



Number of samples per mode as a function of P for fixed N.

Running Times: Serial vs GPU

P	CPU Serial (min)	GPU GTX280 (secs)	Speedup
8192	16.6	2	430
32768	66.7	8	527
131072	270.4	28	572

Running times for Parallel Tempering

- GPU allow huge computational savings for this highly parallelizable method (Lee et al., JCGS 2010).

Monte Carlo methods

Sequential Monte Carlo

A. Doucet

Carcans

Sept. 2011

Generic Problem

- Consider a sequence of probability distributions $\{\pi_n\}_{n\geq 1}$ defined on a sequence of measurable spaces $\{(E_n, \mathcal{F}_n)\}_{n\geq 1}$ where $E_1 = E$, $\mathcal{F}_1 = \mathcal{F}$ and $E_n = E_{n-1} \times E$, $\mathcal{F}_n = \mathcal{F}_{n-1} \times \mathcal{F}$.
- Each distribution $\pi_n(dx_{1:n}) = \pi_n(x_{1:n})\, dx_{1:n}$ is assumed known *up to a normalizing constant*, i.e.

$$\pi_n(x_{1:n}) = \frac{\gamma_n(x_{1:n})}{Z_n}$$

  where $\gamma_n : E_n \rightarrow \mathbb{R}^+$ can be computed pointwise but Z_n cannot.
- We want to estimate expectations of test functions $\varphi_n : E_n \rightarrow \mathbb{R}$

$$\mathbb{E}_{\pi_n}(\varphi_n) = \int \varphi_n(x_{1:n})\, \pi_n(dx_{1:n})$$

  and the normalizing constants Z_n.
- We want to do this **sequentially**; i.e. first π_1 and/or Z_1 at time 1 then π_2 and/or Z_2 at time 2 and so on.

Interacting Particle Methods

- Numerical methods are required.
- We could use Markov chain Monte Carlo (MCMC) to sample from $\{\pi_n\}_{n\geq 1}$ but it is slow & it does not provide directly estimates of $\{Z_n\}_{n\geq 1}$.
- Interacting particle methods aka Sequential Monte Carlo (SMC) are a non-iterative alternative class of methods.
- *Key idea*: if π_{n-1} does not differ too much from π_n then we should be able to reuse our estimate of π_{n-1} to approximate π_n.

Applications

- Inference in non-linear non-Gaussian dynamic models.
- Bayesian inference for complex statistical models.
- Counting problems.
- Rare event simulation.
- Eigenvalue computation.

State-Space Models

- $\{X_n\}_{n\geq 1}$ latent/hidden Markov process with
$$X_1 \sim \mu(\cdot) \text{ and } X_n|(X_{n-1}=x) \sim f(\cdot|x).$$
- $\{Y_n\}_{n\geq 1}$ observation process such that observations are conditionally independent given $\{X_n\}_{n\geq 1}$ and
$$Y_n|(X_n=x) \sim g(\cdot|x).$$
- Very popular class of time series models also known as hidden Markov models.
- **Objective**: Infer the latent process given the observation process.

Examples

- *Linear Gaussian state-space model*

$$\begin{aligned} X_1 &\sim \mathcal{N}(m_1, \Sigma_1), \; X_n = AX_{n-1} + BV_n, \\ Y_n &= CX_n + DW_n \end{aligned}$$

  where $V_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_v)$, $W_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_w)$.

- *Stochastic volatility model*

$$\begin{aligned} X_1 &\sim \mathcal{N}\left(0, \frac{\sigma^2}{1-\alpha^2}\right), \; X_n = \alpha X_{n-1} + V_n, \\ Y_n &= \beta \exp(X_n/2) W_n \end{aligned}$$

  where $|\alpha| < 1$, $V_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, $W_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$.

Inference in State-Space Models

- At time n, we have access to the observations $y_{1:n}$ and are interested in computing

$$p\left(x_{1:n} \middle| y_{1:n} \right) = \frac{p\left(x_{1:n}, y_{1:n} \right)}{p\left(y_{1:n} \right)}$$

and the (marginal) likelihood $p\left(y_{1:n} \right)$ where

$$\begin{aligned} p\left(x_{1:n}, y_{1:n} \right) &= \mu\left(x_1 \right) \prod_{k=2}^{n} f\left(x_k \middle| x_{k-1} \right) \prod_{k=1}^{n} g\left(y_k \middle| x_k \right), \\ p\left(y_{1:n} \right) &= \int \cdots \int p\left(x_{1:n}, y_{1:n} \right) dx_{1:n}. \end{aligned}$$

- In our framework,

$$\pi_n\left(x_{1:n} \right) = p\left(x_{1:n} \middle| y_{1:n} \right), \; \gamma_n\left(x_{1:n} \right) = p\left(x_{1:n}, y_{1:n} \right), \; Z_n = p\left(y_{1:n} \right).$$

Generic Sequence of Target Distributions

- Consider the case where all the target distributions $\{\pi_n\}_{n\geq 1}$ are defined on the same space $E_n = E$.
- *Examples*
  - $\pi_n(x) \propto p(x)\left[p(y|x)\right]^{\phi_n}$ where $\phi_1 = 0 \leq \phi_2 \leq \cdots \leq \phi_P = 1$ (e.g. annealing)
  - $\pi_n(x) = p(x|y_{1:n})$ (sequential Bayesian estimation)
- In these scenarios, MCMC are the standard tools and standard SMC do not apply as they require $E_n = E^n$.

Enlarging Artificially the State-Space

- Consider a new sequence of *artificial* distributions $\left\{\widetilde{\pi}_n\right\}_{n\geq 1}$ defined on $E_n = E^n$ such that

$$\int \widetilde{\pi}_n \left(x_{1:n-1}, x_n\right) dx_{1:n-1} = \pi_n \left(x_n\right).$$

- It is easy to build distribution satisfying this requirement;

$$\widetilde{\pi}_n \left(x_{1:n-1}, x_n\right) = \pi_n \left(x_n\right) \widetilde{\pi}_n \left(x_{1:n-1} \middle| x_n \right)$$

  where $\widetilde{\pi}_n \left(x_{1:n-1} \middle| x_n \right)$ is *any* conditional distribution on E_{n-1}. How to select $\widetilde{\pi}_n$ will be discussed later.

- This allows us to use "standard" SMC and has become an increasingly popular alternative to MCMC.

Using Monte Carlo Methods

- **Problem 1**: $\{\pi_n(x_{1:n})\}_{n\geq 1}$ are typically high dimensional non-standard distributions and we cannot sample from them exactly.
    - A standard approach to sample from high dimensional distributions consists of using MCMC but this is not quite appropriate in our context.
- **Problem 2:** Even if we could sample exactly from $\{\pi_n(x_{1:n})\}_{n\geq 1}$, then the computational complexity of the algorithm would most likely increase with n but we favour algorithms of fixed computational complexity at each time step.
- Importance sampling will allow us to bypass *partially* these problems.

Importance Sampling

- **Importance Sampling (IS)**. For any pdf $q\left(x\right)$ such that $\pi\left(x\right)>0\Rightarrow q\left(x\right)>0$

$$\pi\left(x\right)=\frac{w\left(x\right)q\left(x\right)}{\int w\left(x'\right)q\left(x'\right)dx'}\text{ where }w\left(x\right)=\frac{\gamma\left(x\right)}{q\left(x\right)}$$

where q is called *importance density* and w *importance weight.*

- q can be chosen arbitrarily, in particular easy to sample from

$$X^{(i)}\overset{\text{i.i.d.}}{\sim}q\left(\cdot\right)\Rightarrow\widehat{q}\left(dx\right)=\frac{1}{N}\sum_{i=1}^{N}\delta_{X^{(i)}}\left(dx\right)$$

- Plugging this expression in IS identity

$$\begin{aligned}\widehat{\pi}(dx) &= \frac{w(x)\,\widehat{q}(dx)}{\int w(x')\,\widehat{q}(dx')} = \frac{N^{-1}\sum_{i=1}^{N} w\left(X^{(i)}\right)\delta_{X^{(i)}}(dx)}{N^{-1}\sum_{i=1}^{N} w\left(X^{(i)}\right)} \\ &= \sum_{i=1}^{N} W^{(i)}\delta_{X^{(i)}}(dx)\end{aligned}$$

where

$$W^{(i)} \propto w\left(X^{(i)}\right) \text{ and } \sum_{i=1}^{N} W^{(i)} = 1.$$

- $\pi(x)$ now approximated by weighted sum of delta-masses ⇒ Weights compensate for discrepancy between π and q.

- Now we can approximate $\mathbb{E}_{\pi}[\varphi]$ by

$$\mathbb{E}_{\widehat{\pi}}[\varphi] = \int \varphi(x)\,\widehat{\pi}(dx) = \sum_{i=1}^{N} W^{(i)} \varphi\left(X^{(i)}\right).$$

- We have for $N \gg 1$

$$\begin{aligned} \mathbb{E}\left[\mathbb{E}_{\widehat{\pi}}[\varphi]\right] &\approx \mathbb{E}_{\pi}[\varphi] - N^{-1}\mathbb{E}_{\pi}\left[W(X)\left(\varphi(X) - \mathbb{E}_{\pi}[\varphi]\right)\right], \\ \mathbb{V}\left[\mathbb{E}_{\widehat{\pi}}[\varphi]\right] &\approx N^{-1}\mathbb{E}_{\pi}\left[W(X)\left(\varphi(X) - \mathbb{E}_{\pi}[\varphi]\right)^{2}\right]. \end{aligned}$$

- Estimate of normalizing constant

$$\widehat{Z} = \int \frac{\gamma(x)}{q(x)}\widehat{q}(dx) = \frac{1}{N}\sum_{i=1}^{N} \frac{\gamma\left(X^{(i)}\right)}{q\left(X^{(i)}\right)}$$

and $\mathbb{E}_{q}\left[\widehat{Z}\right] = Z$, $\mathbb{V}_{q}\left[\widehat{Z}\right]/Z^{2} = N^{-1}\left(\mathbb{E}_{q}\left[\left(\frac{\pi(X)}{q(X)} - 1\right)^{2}\right]\right).$

Practical recommendations

- Select q as close to π as possible.
- The variance of the weights is bounded if and only if

$$\int \frac{\gamma^2(x)}{q(x)} dx < \infty.$$

- In practice, try to ensure

$$w(x) = \frac{\gamma(x)}{q(x)} < \infty.$$

Note that in this case, rejection sampling could be used to sample from $\pi(x)$.

Example

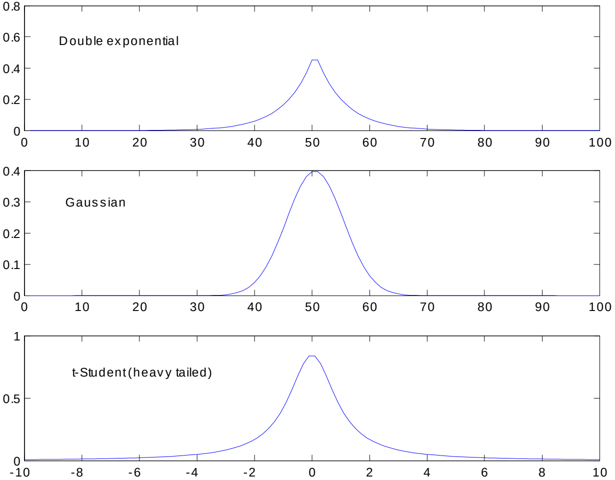


Figure: Target double exponential distributions and two IS distributions

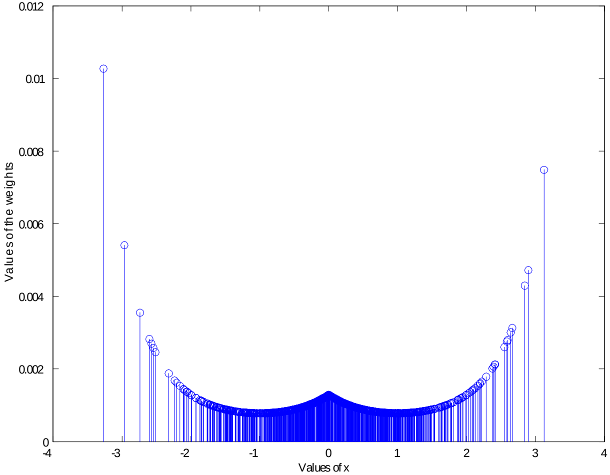


Figure: IS approximation obtained using a Gaussian IS distribution

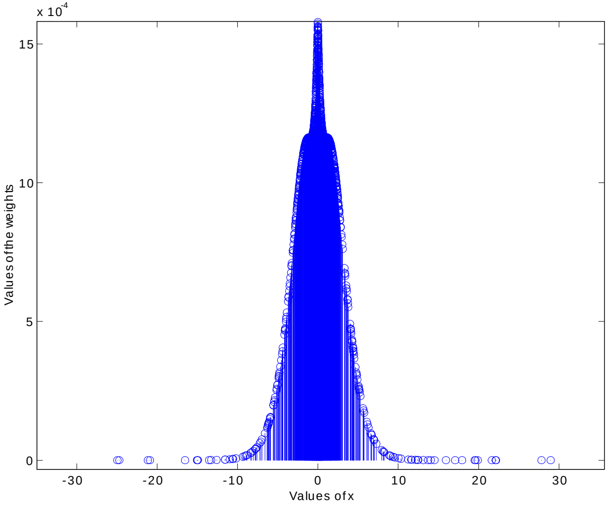


Figure: IS approximation obtained using a Student-t IS distribution

Limitations of Importance Sampling

- MCMC have become prominent in computational statistics as IS is known to scale poorly.
- Consider the case where the target is defined on $\mathbb{R}^n$ and

$$\pi\left(x_{1:n}\right)=\prod_{n=1}^{n}\mathcal{N}\left(x_k;0,1\right),$$

$$\gamma\left(x_{1:n}\right)=\prod_{k=1}^{n}\exp\left(-\frac{x_k^2}{2}\right),\ Z=\left(2\pi\right)^{n/2}.$$

- We select an importance distribution

$$q\left(x_{1:n}\right)=\prod_{k=1}^{n}\mathcal{N}\left(x_k;0,\sigma^2\right).$$

- In this case, we have $\mathbb{V}\left[\widehat{Z}\right]<\infty$ only for $\sigma^2>\frac{1}{2}$ and

$$\frac{\mathbb{V}_{\text{IS}}\left[\widehat{Z}\right]}{Z^2}=\frac{1}{N}\left[\left(\frac{\sigma^4}{2\sigma^2-1}\right)^{n/2}-1\right].$$

Curse of Dimensionality

- The variance increases exponentially with n for any $\frac{1}{2} < \sigma^2 \neq 1$.
- For example, if we select $\sigma^2 = 1.2$ then we have a reasonably good importance distribution as $q(x_k) \approx \pi(x_k)$ but $N \frac{\mathbb{V}_{\text{IS}}[\widehat{Z}]}{Z^2} \approx (1.103)^{n/2}$ which is approximately equal to 1.9×10^{21} for $n = 1000$.
- We would need to use $N \approx 2 \times 10^{23}$ particles to obtain a relative variance $\frac{\mathbb{V}_{\text{IS}}[\widehat{Z}]}{Z^2} = 0.01$.
- Despite this obvious limitations, we will keep on using IS for the time being.

Sequential Importance Sampling (SIS)

- **Aim**: Design an IS method to approximate sequentially $\{\pi_n\}_{n\geq 1}$ and to compute $\{Z_n\}_{n\geq 1}$.
- At time 1, assume we approximate $\pi_1(x_1)$ and Z_1 using an importance density $q_1(x_1)$; that is

$$\begin{aligned} \widehat{\pi}_1(dx_1) &= \sum_{i=1}^{N} W_1^{(i)} \delta_{X_1^{(i)}}(dx) \text{ where } W_1^{(i)} \propto w_1\left(X_1^{(i)}\right), \\ \widehat{Z}_1 &= \frac{1}{N}\sum_{i=1}^{N} w_1\left(X_1^{(i)}\right) \end{aligned}$$

with

$$w_1(x_1) = \frac{\gamma_1(x_1)}{q_1(x_1)}.$$

Building Up the IS Approximation

- At time 2, we want to approximate $\pi_2\left(x_{1:2}\right)$ and Z_2 using an importance density $q_2\left(x_{1:2}\right)$.
- We want to reuse the samples $\left\{X_1^{(i)}\right\}$ from $q_1\left(x_1\right)$ use to build the IS approximation of $\pi_1\left(x_1\right)$. This only makes sense if $\pi_2\left(x_1\right)\approx\pi_1\left(x_1\right)$.
- We select

$$q_2\left(x_{1:2}\right)=q_1\left(x_1\right)q_2\left(x_2\right|x_1\left.\right)$$

so that to obtain $X_{1:2}^{(i)}\sim q_2\left(\cdot\right)$ we only need to sample $X_2^{(i)}\sim q_2\left(\cdot\middle|X_1^{(i)}\right)$.

Updating the IS approximation

- We have to compute the weights

$$\begin{aligned} w_2(x_{1:2}) &= \frac{\gamma_2(x_{1:2})}{q_2(x_{1:2})} = \frac{\gamma_2(x_{1:2})}{q_1(x_1)\, q_2(x_2|\, x_1)} \\ &= \frac{\gamma_1(x_1)}{q_1(x_1)} \frac{\gamma_2(x_{1:2})}{\gamma_1(x_1)\, q_2(x_2|\, x_1)} \\ &= \underbrace{w_1(x_1)}_{\text{previous weight}} \underbrace{\frac{\gamma_2(x_{1:2})}{\gamma_1(x_1)\, q_2(x_2|\, x_1)}}_{\text{incremental weight}} \end{aligned}$$

- For the normalized weights

$$W_2^{(i)} \propto W_1^{(i)} \frac{\gamma_2\left(X_{1:2}^{(i)}\right)}{\gamma_1\left(X_1^{(i)}\right) q_2\left(X_2^{(i)} \middle|\, X_1^{(i)}\right)}$$

Sequential Importance Sampling

- Generally speaking, we use at time n

$$\begin{aligned} q_n(x_{1:n}) &= q_{n-1}(x_{1:n-1})\, q_n(x_n | x_{1:n-1}) \\ &= q_1(x_1)\, q_2(x_2 | x_1) \cdots q_n(x_n | x_{1:n-1}) \end{aligned}$$

  so if $X_{1:n-1}^{(i)} \sim q_{n-1}(\cdot)$ then we only need to sample $X_n^{(i)} \sim q_n\left(\cdot | X_{1:n-1}^{(i)} \right)$.

- The importance weights are updated according to

$$w_n(x_{1:n}) = \frac{\gamma_n(x_{1:n})}{q_n(x_{1:n})} = w_{n-1}(x_{1:n-1}) \frac{\gamma_n(x_{1:n})}{\gamma_{n-1}(x_{1:n-1})\, q_n(x_n | x_{1:n-1})}$$

Sequential Importance Sampling

- At time $n = 1$, sample $X_1^{(i)} \sim q_1(\cdot)$ and set $w_1\left(X_1^{(i)}\right) = \frac{\gamma_1\left(X_1^{(i)}\right)}{q_1\left(X_1^{(i)}\right)}$.
- At time $n \geq 2$
  - sample $X_n^{(i)} \sim q_n\left(\cdot \middle| X_{1:n-1}^{(i)}\right)$
  - compute $w_n\left(X_{1:n}^{(i)}\right) = w_{n-1}\left(X_{1:n-1}^{(i)}\right) \frac{\gamma_n\left(X_{1:n}^{(i)}\right)}{\gamma_{n-1}\left(X_{1:n-1}^{(i)}\right) q_n\left(X_n^{(i)} \middle| X_{1:n-1}^{(i)}\right)}$.
- At any time n, we have

$$X_{1:n}^{(i)} \sim q_n(\cdot), \quad w_n\left(X_{1:n}^{(i)}\right) = \frac{\gamma_n\left(X_{1:n}^{(i)}\right)}{q_n\left(X_{1:n}^{(i)}\right)}$$

thus we can obtain easily an IS approximation of $\pi_n(x_{1:n})$ and of Z_n.

Application to Stochastic Volatility Model

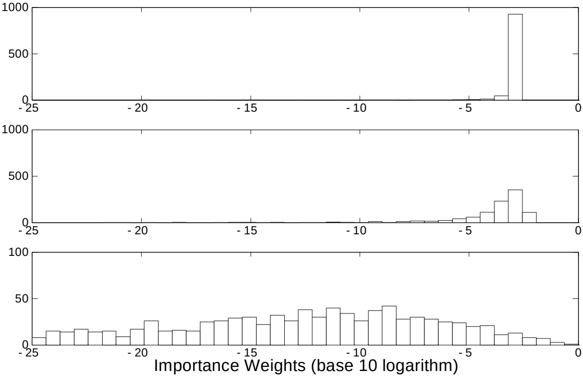


Figure: Histograms of $\log_{10}\left(W_n^{(i)}\right)$ for $n = 1$ (top), $n = 50$ (middle) and $n = 100$ (bottom).

- The algorithm performance collapse as n increases as expected.

Locally Optimal Importance Density

- One sensible strategy consists of selecting $q_n(x_n | x_{1:n-1})$ at time n so as to minimize the variance of the importance weights.
- We have for the importance weight

$$\begin{aligned} w_n(x_{1:n}) &= \frac{\gamma_n(x_{1:n})}{q_{n-1}(x_{1:n-1})\, q_n(x_n | x_{1:n-1})} \\ &= \frac{Z_n \pi_n(x_{1:n-1})}{q_{n-1}(x_{1:n-1})} \frac{\pi_n(x_n | x_{1:n-1})}{q_n(x_n | x_{1:n-1})} \end{aligned}$$

- It follows directly that we have

$$q_n^{\text{opt}}(x_n | x_{1:n-1}) = \pi_n(x_n | x_{1:n-1})$$

and

$$\begin{aligned} w_n(x_{1:n}) &= w_{n-1}(x_{1:n-1}) \frac{\gamma_n(x_{1:n})}{\gamma_{n-1}(x_{1:n-1})\, \pi_n(x_n | x_{1:n-1})} \\ &= w_{n-1}(x_{1:n-1}) \frac{\gamma_n(x_{1:n-1})}{\gamma_{n-1}(x_{1:n-1})} \end{aligned}$$

Application to State-Space Models

- In the case of state-space models, we have

$$q_n^{\text{opt}}\left(x_n \mid x_{1:n-1}\right) = p\left(x_n \mid y_n, x_{n-1}\right) = \frac{g\left(y_n \mid x_n\right) f\left(x_n \mid x_{n-1}\right)}{p\left(y_n \mid x_{n-1}\right)}$$

- In this case,

$$\begin{aligned} w_n\left(x_{1:n}\right) &= w_{n-1}\left(x_{1:n-1}\right) \frac{p\left(x_{1:n}, y_{1:n}\right)}{p\left(x_{1:n-1}, y_{1:n-1}\right) p\left(x_n \mid y_n, x_{n-1}\right)} \\ &= w_{n-1}\left(x_{1:n-1}\right) p\left(y_n \mid x_{n-1}\right). \end{aligned}$$

- Whenever $p\left(x_n \mid y_n, x_{n-1}\right)$ is not easy to sample and/or $p\left(y_n \mid x_{n-1}\right)$ cannot be computed, you can use sthe EKF, UKF or any standard deterministic approximation to approximate $p\left(x_n \mid y_n, x_{n-1}\right)$.

Resampling

- *Intuitive KEY idea*: As the time index n increases, the variance of the unnormalized weights $\left\{ w_n \left(X_{1:n}^{(i)} \right) \right\}$ tends to increase and all the mass is concentrated on a few random samples/particles. We propose to reset the approximation by getting rid in a principled way of the particles with low weights $W_n^{(i)}$ (relative to $1/N$) and multiply the particles with high weights $W_n^{(i)}$ (relative to $1/N$).
- The main reason is that if a particle at time n has a low weight then typically it will still have a low weight at time $n+1$ (though it is easy to come up with counterexamples).
- You want to focus your computational efforts on the "promising" parts of the space.

Multinomial Resampling

- At time n, SIS provides the following approximation of the target $\pi_n\left(dx_{1:n}\right)$

$$\widehat{\pi}_n\left(dx_{1:n}\right)=\sum_{i=1}^{N}W_n^{(i)}\delta_{X_{1:n}^{(i)}}\left(dx_{1:n}\right).$$

- The simplest resampling scheme consists of sampling N times $\widetilde{X}_{1:n}^{(i)}\sim\widehat{\pi}_n\left(\cdot\right)$ to build the new approximation

$$\widetilde{\pi}_n\left(dx_{1:n}\right)=\frac{1}{N}\sum_{i=1}^{N}\delta_{\widetilde{X}_{1:n}^{(i)}}\left(dx_{1:n}\right).$$

- The new resampled particles $\left\{\widetilde{X}_{1:n}^{(i)}\right\}$ are *approximately* distributed according to π_n but now statistically dependent.

- Note that we can rewrite
$$\widetilde{\pi}_n\left(dx_{1:n}\right)=\sum_{i=1}^{N}\frac{N_n^{(i)}}{N}\delta_{X_{1:n}^{(i)}}\left(dx_{1:n}\right)$$
where $\left(N_n^{(1)},...,N_n^{(N)}\right)\sim\mathcal{M}\left(N;W_n^{(1)},...,W_n^{(N)}\right)$ thus $\mathbb{E}\left[N_n^{(i)}\right]=NW_n^{(i)}$, $\mathbb{V}\left[N_n^{(1)}\right]=NW_n^{(i)}\left(1-W_n^{(i)}\right)$.
- The resampling step is an unbiased operation
$$\mathbb{E}\left[\widetilde{\pi}_n\left(dx_{1:n}\right)\middle|\,\widehat{\pi}_n\left(dx_{1:n}\right)\right]=\widehat{\pi}_n\left(dx_{1:n}\right)$$
but clearly it introduces some errors “locally” in time. That is for any test function, we have
$$\mathbb{V}_{\widetilde{\pi}_n}\left[\varphi\left(X_{1:n}\right)\right]\geq\mathbb{V}_{\widehat{\pi}_n}\left[\varphi\left(X_{1:n}\right)\right]$$
- Resampling can be beneficial for future time steps (sometimes).
- Better resampling steps can be designed such that $\mathbb{E}\left[N_n^{(i)}\right]=NW_n^{(i)}$ but $\mathbb{V}\left[N_n^{(i)}\right]<NW_n^{(i)}\left(1-W_n^{(i)}\right)$; residual resampling, minimal entropy resampling etc. (Cappé et al., 2005).

Sequential Importance Sampling Resampling

- At time $n = 1$, sample $X_1^{(i)} \sim q_1(\cdot)$ and set $w_1\left(X_1^{(i)}\right) = \frac{\gamma_1\left(X_1^{(i)}\right)}{q_1\left(X_1^{(i)}\right)}$.
- Resample $\left\{X_1^{(i)}, W_1^{(i)}\right\}$ to obtain new particles also denoted $\left\{X_1^{(i)}\right\}$
- At time $n \geq 2$
  - sample $X_n^{(i)} \sim q_n\left(\cdot | X_{1:n-1}^{(i)}\right)$
  - compute $w_n\left(X_{1:n}^{(i)}\right) = \frac{\gamma_n\left(X_{1:n}^{(i)}\right)}{\gamma_{n-1}\left(X_{1:n-1}^{(i)}\right) q_n\left(X_n^{(i)} \middle| X_{1:n-1}^{(i)}\right)}$.
- Resample $\left\{X_{1:n}^{(i)}, W_n^{(i)}\right\}$ to obtain new particles also denoted $\left\{X_{1:n}^{(i)}\right\}$

SMC Estimates

- At any time n, we have two approximations of $\pi_n\left(dx_{1:n}\right)$

$$
\begin{aligned}
\widehat{\pi}_n\left(dx_{1:n}\right) &= \sum_{i=1}^{N} W_n^{(i)} \delta_{X_{1:n}^{(i)}}\left(dx_{1:n}\right) \text{ (before resampling)} \\
\widetilde{\pi}_n\left(dx_{1:n}\right) &= \frac{1}{N} \sum_{i=1}^{N} \delta_{X_{1:n}^{(i)}}\left(dx_{1:n}\right) \text{ (after resampling).}
\end{aligned}
$$

- We also have

$$\frac{Z_n}{Z_{n-1}} = \int w_n\left(x_{1:n}\right) \pi_{n-1}\left(x_{1:n-1}\right) q_n\left(\left. x_n \right| x_{1:n-1}\right) dx_{1:n}$$

so an estimate is given by

$$\widehat{\frac{Z_n}{Z_{n-1}}} = \frac{1}{N} \sum_{i=1}^{N} w_n\left(X_{1:n}^{(i)}\right).$$

Unbiasedness of the Normalizing Constant Estimate

- Let
$$\widehat{Z_n} = \widehat{Z}_1 \prod_{k=2}^{n} \frac{\widehat{Z_k}}{Z_{k-1}} = \prod_{k=1}^{n} \left(\frac{1}{N} \sum_{i=1}^{N} w_k \left(X_{1:k}^{(i)} \right) \right)$$
- As long as the resampling scheme used in unbiased; i.e. $\mathbb{E}\left[N_n^{(i)} \right] = N W_n^{(i)}$ then
$$\mathbb{E}\left(\widehat{Z_n} \right) = Z_n$$
as in the standard SIS case.
- This remarkable properties will be exploited later on in the context of particle MCMC algorithms.

Sequential Monte Carlo for Hidden Markov Models

- At time $n=1$, sample $X_1^{(i)} \sim q_1(\cdot)$ and set $w_1\left(X_1^{(i)}\right) = \frac{\mu\left(X_1^{(i)}\right) g\left(y_1 | X_1^{(i)}\right)}{q\left(X_1^{(i)} \middle| y_1\right)}$.
- Resample $\left\{X_1^{(i)}, W_1^{(i)}\right\}$ to obtain new particles also denoted $\left\{X_1^{(i)}\right\}$
- At time $n \geq 2$
  - sample $X_n^{(i)} \sim q\left(\cdot | \, y_n, X_{n-1}^{(i)}\right)$
  - compute $w_n\left(X_{1:n}^{(i)}\right) = \frac{f\left(X_n^{(i)} \middle| X_{n-1}^{(i)}\right) g\left(y_n | X_n^{(i)}\right)}{q\left(X_n^{(i)} \middle| y_n, X_{n-1}^{(i)}\right)}$.
- Resample $\left\{X_{1:n}^{(i)}, W_n^{(i)}\right\}$ to obtain new particles also denoted $\left\{X_{1:n}^{(i)}\right\}$

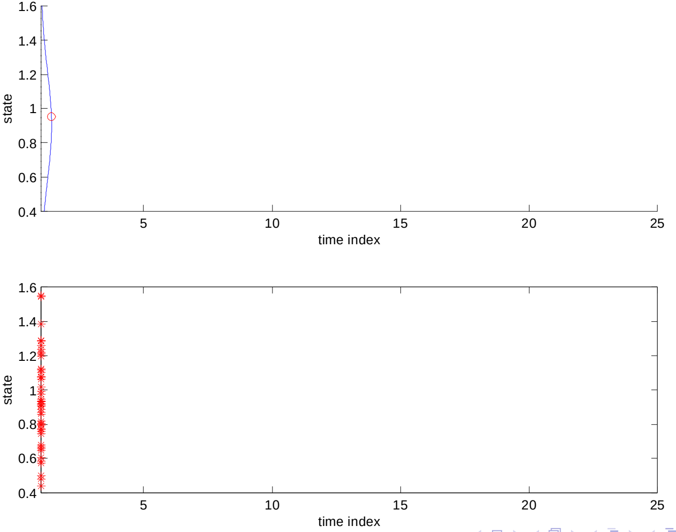
- **Example**: Linear Gaussian model

$$\begin{aligned} X_1 &\sim \mathcal{N}(0,1), \ X_n = \alpha X_{n-1} + \sigma_v V_n, \\ Y_n &= X_n + \sigma_w W_n \end{aligned}$$

  where $V_n \sim \mathcal{N}(0,1)$ and $W_n \sim \mathcal{N}(0,1)$.

- We know that $p(x_{1:n}|y_{1:n})$ is Gaussian and its parameters can be computed using Kalman techniques. In particular $p(x_n|y_{1:n})$ is also a Gaussian whose parameters can be computed using the Kalman filter.
- We apply the SMC method with $q(x_n|y_n, x_{n-1}) = f(x_n|x_{n-1}) = \mathcal{N}(x_n; \alpha x_{n-1}, \sigma_v^2)$.

Illustration of the Degeneracy Problem (Figures by Olivier Cappé)



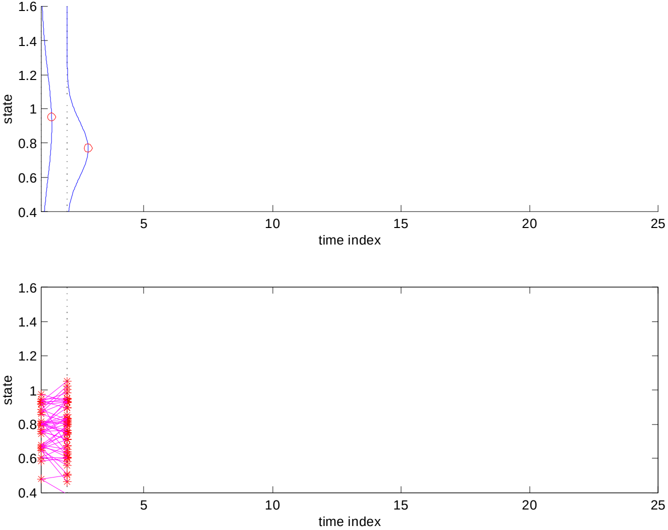


Figure: $p\left(x_1 \mid y_1\right)$, $p\left(x_2 \mid y_{1:2}\right)$and $\widehat{\mathbb{E}}\left[X_1 \mid y_1\right]$, $\widehat{\mathbb{E}}\left[X_2 \mid y_{1:2}\right]$ (top) and particle approximation of $p\left(x_{1:2} \mid y_{1:2}\right)$ (bottom)

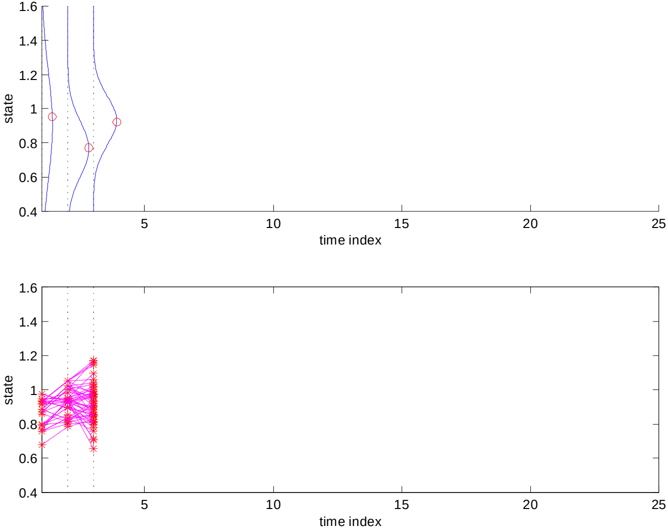


Figure: $p\left(x_k \mid y_{1:k}\right)$ and $\widehat{\mathbb{E}}\left[X_k \mid y_{1:k}\right]$ for $k = 1, 2, 3$ (top) and particle approximation of $p\left(x_{1:3} \mid y_{1:3}\right)$ (bottom)

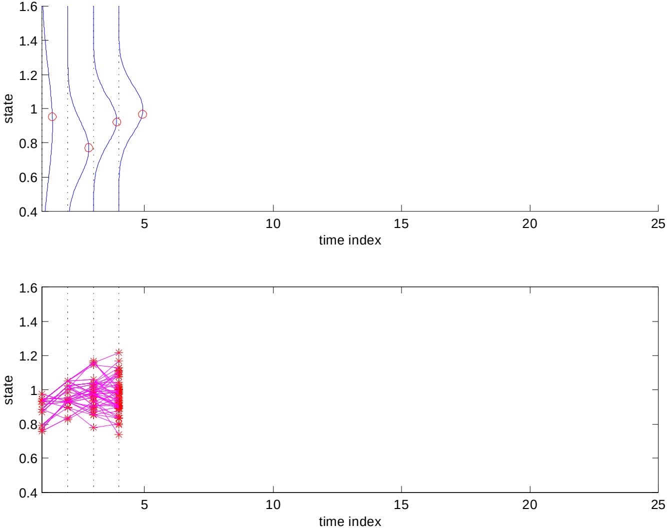


Figure: $p\left(x_k \mid y_{1:k}\right)$ and $\widehat{\mathbb{E}}\left[X_k \mid y_{1:k}\right]$ for $k = 1, .., 4$ (top) and particle approximation of $p\left(x_{1:4} \mid y_{1:4}\right)$ (bottom)

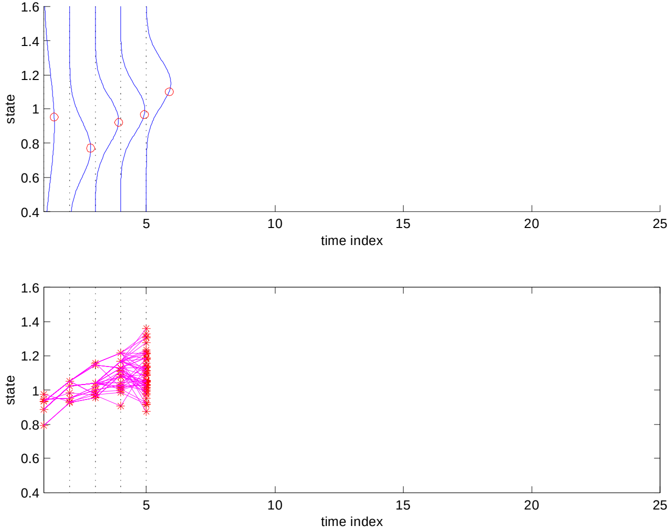


Figure: $p\left(x_k \mid y_{1:k}\right)$ and $\widehat{\mathbb{E}}\left[X_k \mid y_{1:k}\right]$ for $k = 1, ..., 5$ (top) and particle approximation of $p\left(x_{1:5} \mid y_{1:5}\right)$ (bottom)

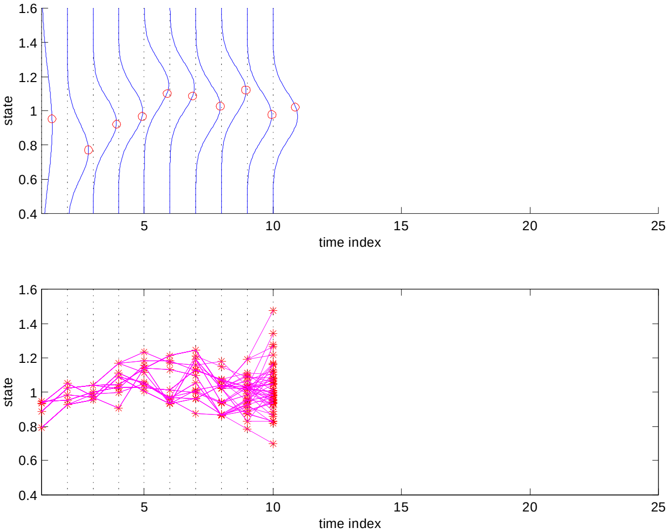


Figure: $p\left(x_k \middle| y_{1:k}\right)$ and $\widehat{\mathbb{E}}\left[X_k \middle| y_{1:k}\right]$ for $k = 1, ..., 10$ (top) and particle approximation of $p\left(x_{1:10} \middle| y_{1:10}\right)$ (bottom)

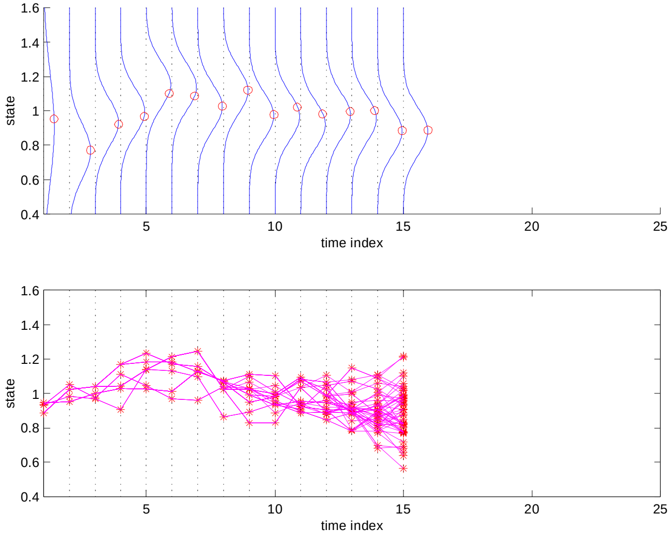


Figure: $p\left(x_k \middle| y_{1:k}\right)$ and $\widehat{\mathbb{E}}\left[X_k \middle| y_{1:k}\right]$ for $k = 1, ..., 15$ (top) and particle approximation of $p\left(x_{1:15} \middle| y_{1:15}\right)$ (bottom)

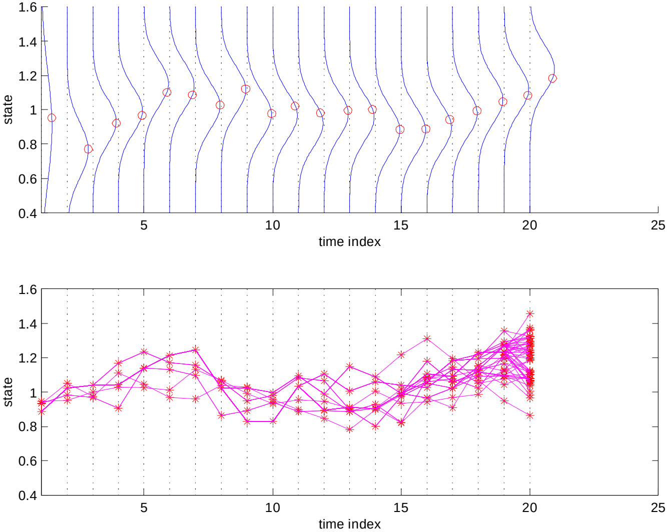


Figure: $p\left(x_k \mid y_{1:k}\right)$ and $\widehat{\mathbb{E}}\left[X_k \mid y_{1:k}\right]$ for $k = 1, ..., 20$ (top) and particle approximation of $p\left(x_{1:20} \mid y_{1:20}\right)$ (bottom)

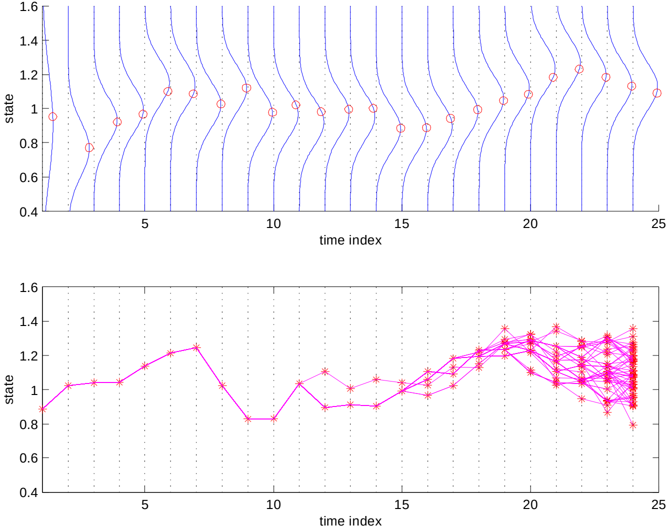


Figure: $p\left(x_k \mid y_{1:k}\right)$ and $\widehat{\mathbb{E}}\left[X_k \mid y_{1:k}\right]$ for $k = 1, ..., 24$ (top) and particle approximation of $p\left(x_{1:24} \mid y_{1:24}\right)$ (bottom)

Remarks

- Empirically this SMC strategy performs well in terms of estimation of the marginals $\left\{ p\left(x_n \middle| y_{1:n} \right) \right\}_{n \geq 1}$. This is what is only necessary in many applications thankfully.
- However, the joint distribution $p\left(x_{1:k} \middle| y_{1:k} \right)$ is poorly estimated when k is large; i.e. we have in the previous example

$$\widehat{p}\left(x_{1:11} \middle| y_{1:24} \right) = \delta_{X_{1:11}^*}\left(x_{1:11} \right).$$

- **Degeneracy problem**. For any N and any k, there exists $n\left(k, N\right)$ such that for any $n \geq n\left(k, N\right)$

$$\widehat{p}\left(dx_{1:k} \middle| y_{1:n} \right) = \delta_{X_{1:k}^*}\left(dx_{1:k} \right);$$

  $\widehat{p}\left(dx_{1:n} \middle| y_{1:n} \right)$ is an unreliable approximation of $p\left(dx_{1:n} \middle| y_{1:n} \right)$ as $n \nearrow$.
- **Resampling only partially solves our problem**.

Another Illustration of the Degeneracy Phenomenon

- For the linear Gaussian state-space model described before, we can compute exactly S_n / n where

$$S_n = \int \left(\sum_{k=1}^{n} x_k^2 \right) p\left(dx_{1:n} \middle| y_{1:n} \right)$$

using Kalman techniques.

- We compute the SMC estimate of this quantity using $\widehat{S}_n / n$ where

$$\widehat{S}_n = \int \left(\sum_{k=1}^{n} x_k^2 \right) \widehat{p}\left(dx_{1:n} \middle| y_{1:n} \right)$$

- This estimate can be computed sequentially.

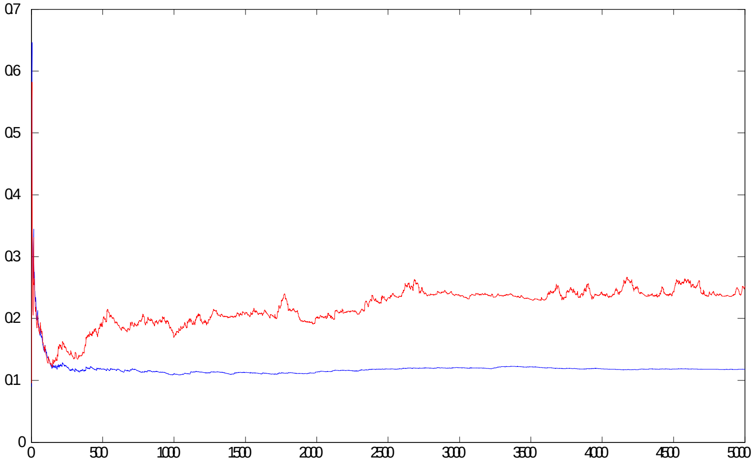


Figure: Sufficient statistics computed exactly through the Kalman smoother (blue) and the SMC method (red).

Some Convergence Results for SMC

- Numerous convergence results for SMC are available; see (Del Moral, 2004): Lp bounds, CLT, concentration inequalities, large deviations.
- In particular we can prove rather easily that for any bounded function φ and any $p \geq 1$ (Del Moral, Crisan & Lyons, 1997)

$$\mathbb{E}\left[\left|\int \varphi_n\left(x_{1:n}\right)\left(\widehat{\pi}_n\left(dx_{1:n}\right)-\pi_n\left(dx_{1:n}\right)\right)\right|^p\right]^{1/p} \leq \frac{c\left(n\right)b\left(p\right)\left\|\varphi\right\|_\infty}{\sqrt{N}}.$$

- It is not a very informative result as $c\left(n\right)$ increases polynomially/exponentially with time.
- To achieve a fixed precision, this would require to use a time-increasing number of particles N. Without any additional assumption, we cannot expect to get better results.

Uniform In Time Convergence Results for SMC

- Under strong mixing assumptions, you can obtain much stronger results

$$\mathbb{E}\left[\left|\int \varphi_n(x_n)\left(\widehat{\pi}_n(dx_n)-\pi_n(dx_n)\right)\right|^p\right]^{1/p} \leq \frac{c_1\ b(p)\|\varphi\|_\infty}{\sqrt{N}}$$

  i.e. there is no accumulation of numerical errors over time for the marginals (Del Moral, 2004).

- Under mixing assumptions, we have (Cérou et al., 2011, Whiteley et al., 2011)

$$\mathbb{E}\left(\left(\frac{\widehat{Z}_n}{Z_n}-1\right)^2\right) \leq \frac{c_2\ n}{N}$$

- Under mixing assumptions, if $\overline{\pi}_n(dx_{1:n}) = \mathbb{E}\left(\widehat{\pi}_n(dx_{1:n})\right)$ then (Del Moral et al., 2010)

$$\|\overline{\pi}_n-\pi_n\|_{\text{tv}} \leq \frac{c_3\ n}{N}$$

Back to our toy example

- Consider the case where the target is defined on $\mathbb{R}^n$ and

$$\pi\left(x_{1:n}\right)=\prod_{n=1}^{n}\mathcal{N}\left(x_k;0,1\right),$$

$$\gamma\left(x_{1:n}\right)=\prod_{k=1}^{n}\exp\left(-\frac{x_k^2}{2}\right),\;\; Z=\left(2\pi\right)^{n/2}.$$

- We select an importance distribution

$$q\left(x_{1:n}\right)=\prod_{k=1}^{n}\mathcal{N}\left(x_k;0,\sigma^2\right).$$

- For SMC, the asymptotic variance is finite only when $\sigma^2 > \frac{1}{2}$ and

$$\frac{\mathbb{V}_{\text{SMC}}\left[\widehat{Z}_n\right]}{Z_n^2} \approx \frac{1}{N}\left[\int \frac{\pi_n^2\left(x_1\right)}{q_1\left(x_1\right)}dx_1 - 1 + \sum_{k=2}^{n}\int \frac{\pi_n^2\left(x_k\right)}{q_k\left(x_k\right)}dx_k - 1\right]$$
$$= \frac{n}{N}\left[\left(\frac{\sigma^4}{2\sigma^2-1}\right)^{1/2} - 1\right]$$

compared to

$$\frac{\mathbb{V}_{\text{IS}}\left[\widehat{Z}_n\right]}{Z_n^2} = \frac{1}{N}\left[\left(\frac{\sigma^4}{2\sigma^2-1}\right)^{n/2} - 1\right]$$

for SIS.

- If select $\sigma^2 = 1.2$ then it is necessary to use $N \approx 2 \times 10^{23}$ particles to obtain $\frac{\mathbb{V}_{\text{IS}}\left[\widehat{Z}_n\right]}{Z_n^2} = 10^{-2}$ for $n = 1000$.
- To obtain $\frac{\mathbb{V}_{\text{SMC}}\left[\widehat{Z}_n\right]}{Z_n^2} = 10^{-2}$, SMC requires only $N \approx 10^4$ particles: an improvement by 19 orders of magnitude.

Fighting Degeneracy Using MCMC Steps

- A standard way to limit degeneracy is known as the Resample-Move algorithm (Gilks & Berzuini, 2001). It relies upon MCMC kernels as a principled way to "jitter" the particle locations.
- A Markov kernel $K_n\left(\left.x'_{1:n}\right| x_{1:n}\right)$ of invariant distribution $\pi_n\left(x_{1:n}\right)$ is a Markov transition kernel with the property that

$$\int \pi_n\left(x_{1:n}\right) K_n\left(\left.x'_{1:n}\right| x_{1:n}\right) dx_{1:n} = \pi_n\left(x'_{1:n}\right).$$

- *Example.* Set $X'_{1:n-L} = X_{1:n-L}$ then sample X'_{n-L+1} from $\pi_n\left(\left.x_{n-L+1}\right| x'_{1:n-L}, x_{n-L+2:n}\right)$, sample X'_{n-L+2} from $\pi_n\left(\left.x_{n-L+2}\right| x'_{1:n-L+1}, x_{n-L+3:n}\right)$ and so on until we sample X'_n from $\pi_n\left(\left.x_n\right| x'_{1:n-1}\right)$; that is

$$K_n\left(\left.x'_{1:n}\right| x_{1:n}\right) = \delta_{x_{1:n-L}}\left(x'_{1:n-L}\right) \prod_{k=n-L+1}^{n} \pi_n\left(\left.x'_k\right| x'_{1:k-1}, x_{k+1:n}\right).$$

- In the SMC context, we typically do not use ergodic kernels as this would require sampling an increasing number of variables at each time step; i.e. we restrict ourselves to updating the variables X

Summary

- Resampling can drastically improve the performance of SIS in models having 'good' mixing properties; e.g. state-space models: this can be verified experimentally and theoretically.
- Resampling does not solve all our problems; at best only the SMC approximations of the most recent marginals $\pi_n\left(x_{n-L+1:n}\right)$ are reliable; i.e. we can have uniform (in time) convergence bounds.

Online Bayesian Parameter Estimation

- Assume we have

$$X_n | (X_{n-1} = x_{n-1}) \sim f_\theta (x_n | x_{n-1}),$$
$$Y_n | (X_n = x_n) \sim g_\theta (y_n | x_n),$$

where θ is an *unknown* static parameter with prior $p(\theta)$.

- Given data $y_{1:n}$, inference relies on

$$p(\theta, x_{1:n} | y_{1:n}) = p(\theta | y_{1:n}) \, p_\theta (x_{1:n} | y_{1:n})$$

where

$$p(\theta | y_{1:n}) \propto p_\theta (y_{1:n}) \, p(\theta).$$

- SMC methods apply as it is a standard model with extended state $Z_n = (X_n, \theta_n)$ where

$$f(z_n | z_{n-1}) = \underbrace{\delta_{\theta_{n-1}} (\theta_n)}_{\text{practical problems}} f_{\theta_n} (x_n | x_{n-1}), \; g(y_n | z_n) = g_\theta (y_n | x_n).$$

Cautionary Warning

- For fixed θ, $\mathbb{V}\left[\widehat{p}_{\theta}\left(y_{1:n}\right)\right] / p_{\theta}^{2}\left(y_{1:n}\right)$ is in Cn/N.
- In a Bayesian context, the problem is even more complex as $p\left(\theta \mid y_{1:n}\right) \propto p_{\theta}\left(y_{1:n}\right) p\left(\theta\right)$ and we have $\theta_{n}=\theta$ for all n so the latent process does not enjoy mixing properties.
- An attractive idea consists of using MCMC steps on θ; e.g. (Andrieu, De Freitas & D.,1999; Fearnhead, 2002; Gilks & Berzuini 2001; Storvik, 2002; Polson et al., 2010) so as to introduce some "noise" on the θ component of the state.
- When $p\left(\theta \mid y_{1:n}, x_{1:n}\right)=p\left(\theta \mid s_{n}\left(x_{1:n}, y_{1:n}\right)\right)$ where $s_{n}\left(x_{1:n}, y_{1:n}\right)$ is a fixed-dimensional of sufficient statistics, the algorithm is particularly elegant but still implicitly relies on SMC approximation of $p\left(x_{1:n} \mid y_{1:n}\right)$ so degeneracy will creep in.

Example

- Linear Gaussian state-space model

$$X_1 \sim \mathcal{N}\left(0, \sigma_0^2\right) \text{ and } X_k = \theta X_{k-1} + \sigma_V V_k, \;\; V_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$$
$$Y_k = X_k + \sigma_W W_k, \;\; W_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1).$$

- We set $p(\theta) \propto 1_{(-1,1)}(\theta)$ so

$$p\left(\theta \mid y_{1:n}, x_{1:n}\right) \propto \mathcal{N}\left(\theta; m_n, \sigma_n^2\right) 1_{(-1,1)}(\theta)$$

where

$$\sigma_n^2 = S_{2,n}^{-1}, \;\; m_n = S_{2,n}^{-1} S_{1,n}$$

with

$$S_{1,n} = \sum_{k=2}^{n} x_{k-1} x_k, \;\; S_{2,n} = \sum_{k=2}^{n} x_{k-1}^2$$

SMC with MCMC Step for Parameter Estimation

- At time $n-1$, $\left(\theta_{n-1}^{(i)}, X_{n-1}^{(i)}, S_{n-1}^{(i)}\right)$ we have

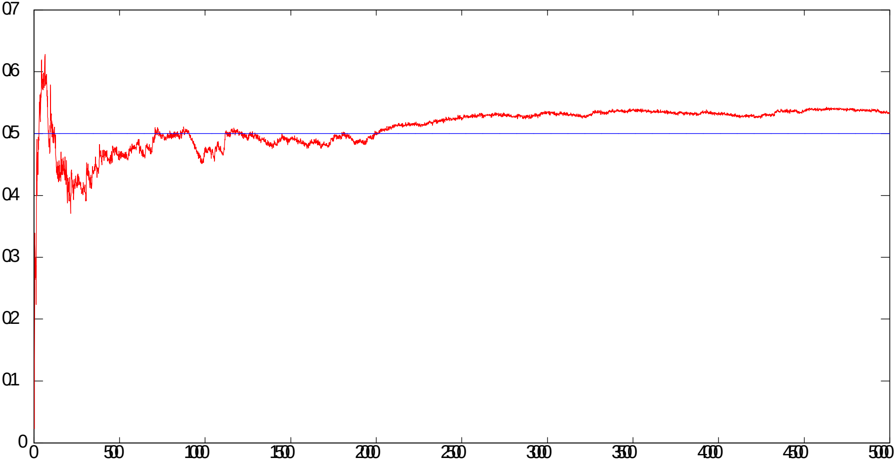
$$\widehat{p}\left(\left. d\theta, dx_{n-1}, ds_{n-1}\right| y_{1:n-1}\right) = \frac{1}{N}\sum_{i=1}^{N}\delta_{\left(\theta_{n-1}^{(i)}, X_{n-1}^{(i)}, S_{n-1}^{(i)}\right)}\left(d\theta, dx_{n-1}, ds_{n-1}\right).$$

- Sample $X_n^{(i)} \sim f_{\theta_{n-1}^{(i)}}\left(\cdot\left| X_{n-1}^{(i)}\right.\right)$, set $S_{1,n}^{(i)} = S_{1,n-1}^{(i)} + X_{n-1}^{(i)}X_n^{(i)}$, $S_{2,n}^{(i)} = S_{2,n-1}^{(i)} + \left(X_{n-1}^{(i)}\right)^2$, $W_n^{(i)} \propto g_{\theta_{n-1}^{(i)}}\left(y_n\left| X_n^{(i)}\right.\right)$ and

$$\widetilde{p}\left(\left. d\theta, dx_n, ds_n\right| y_{1:n}\right) = \sum_{i=1}^{N} W_n^{(i)}\delta_{\left(\theta_{n-1}^{(i)}, \widetilde{X}_n^{(i)}, S_n^{(i)}\right)}\left(d\theta, dx_n, ds_n\right),$$

- Resample $\left(X_n^{(i)}, S_n^{(i)}\right) \sim \widetilde{p}\left(\left. dx_n, ds_n\right| y_{1:n}\right)$ then sample $\theta_n^{(i)} \sim \mathcal{N}\left(\theta; \left(S_{2,n}^{(i)}\right)^{-1} S_{1,n}^{(i)}, \left(S_{2,n}^{(i)}\right)^{-1}\right)1_{(-1,1)}(\theta)$ to obtain $\widehat{p}\left(\left. d\theta, dx_n, ds_n\right| y_{1:n}\right) = \frac{1}{N}\sum_{i=1}^{N}\delta_{\left(\theta_n^{(i)}, X_n^{(i)}, S_n^{(i)}\right)}\left(d\theta, dx_n, ds_n\right)$.

Illustration of the Degeneracy Problem



SMC estimate of $\mathbb{E}\left[\theta \middle| y_{1:n}\right]$, as n increases the degeneracy creeps in.

Online Bayesian Parameter Estimation

- All proposed procedures for online Bayesian parameter estimation are deficient.
- Either some artificial dynamics is introduced but then we cannot expect to approximate $\left\{ p\left(\theta, x_{1:n} \middle| y_{1:n} \right) \right\}_{n \geq 1}$; e.g. (Liu & West, 2001; Flury & Shephard, 2010);
- Methods based on MCMC steps are elegant but do suffer from the degeneracy problem and provide unreliable approximations.

Offline Bayesian Parameter Estimation

- Given a collection of observations $y_{1:T} := (y_1, ..., y_T)$, T being now fixed, we are interested in carrying out inference about θ and $X_{1:T}$.
- Inference relies on the posterior density

$$\begin{aligned} p\left(\theta, x_{1:T} \middle| y_{1:T}\right) &= p\left(\theta \middle| y_{1:T}\right) p_\theta\left(x_{1:T} \middle| y_{1:T}\right) \\ &\propto p\left(\theta, x_{1:T}, y_{1:T}\right) \end{aligned}$$

where

$$p\left(\theta, x_{1:T}, y_{1:T}\right) \propto p\left(\theta\right) \mu_\theta\left(x_1\right) \prod_{n=2}^{T} f_\theta\left(x_n \middle| x_{n-1}\right) \prod_{n=1}^{T} g_\theta\left(y_n \middle| x_n\right) \; .$$

- We show how to address this problem using particle MCMC (Andrieu, D. & Holenstein, *JRSS* B, 2010).

Common MCMC Approaches and Limitations

- **MCMC Idea**: Simulate an ergodic Markov chain $\{\theta(i), X_{1:T}(i)\}_{i\geq 0}$ of invariant distribution $p(\theta, x_{1:T} | y_{1:T})$... infinite number of possibilities.
- Typical strategies consists of updating iteratively $X_{1:T}$ conditional upon θ then θ conditional upon $X_{1:T}$.
- To update $X_{1:T}$ conditional upon θ, use MCMC kernels updating subblocks according to $p_\theta(x_{n:n+K-1} | y_{n:n+K-1}, x_{n-1}, x_{n+K})$.
- Standard MCMC algorithms are inefficient if θ and $X_{1:T}$ are strongly correlated.
- Strategy impossible to implement when it is only possible to sample from the prior but impossible to evaluate it pointwise.

Metropolis-Hastings (MH) Sampling

- To bypass these problems, we want to update jointly θ and $X_{1:T}$.
- Assume that the current state of our Markov chain is $(\theta, x_{1:T})$, we propose to update simultaneously the parameter and the states using a proposal

$$q\left(\left(\theta^{*}, x_{1:T}^{*}\right) \middle| \left(\theta, x_{1:T}\right)\right) = q\left(\theta^{*} \middle| \theta\right) \; q_{\theta^{*}}\left(x_{1:T}^{*} \middle| y_{1:T}\right).$$

- The proposal $\left(\theta^{*}, x_{1:T}^{*}\right)$ is accepted with MH acceptance probability

$$1 \wedge \frac{p\left(\theta^{*}, x_{1:T}^{*} \middle| y_{1:T}\right)}{p\left(\theta, x_{1:T} \middle| y_{1:T}\right)} \frac{q\left(\left(x_{1:T}, \theta\right) \middle| \left(x_{1:T}^{*}, \theta^{*}\right)\right)}{q\left(\left(x_{1:T}^{*}, \theta^{*}\right) \middle| \left(x_{1:T}, \theta\right)\right)}$$

- **Problem**: Designing a proposal $q_{\theta^{*}}\left(x_{1:T}^{*} \middle| y_{1:T}\right)$ such that the acceptance probability is not extremely small is very difficult.

“Idealized” Marginal MH Sampler

- Consider the following so-called marginal Metropolis-Hastings (MH) algorithm which uses as a proposal

$$q\left(\left(x_{1:T}^{*},\theta^{*}\right)\middle|\left(x_{1:T},\theta\right)\right)=q\left(\theta^{*}\middle|\theta\right)p_{\theta^{*}}\left(x_{1:T}^{*}\middle|y_{1:T}\right).$$

- The MH acceptance probability is

$$\begin{aligned}1\wedge\frac{p\left(\theta^{*},x_{1:T}^{*}\middle|y_{1:T}\right)}{p\left(\theta,x_{1:T}\middle|y_{1:T}\right)}\frac{q\left(\left(x_{1:T},\theta\right)\middle|\left(x_{1:T}^{*},\theta^{*}\right)\right)}{q\left(\left(x_{1:T}^{*},\theta^{*}\right)\middle|\left(x_{1:T},\theta\right)\right)}\\=1\wedge\frac{p_{\theta^{*}}\left(y_{1:T}\right)p\left(\theta^{*}\right)}{p_{\theta}\left(y_{1:T}\right)p\left(\theta\right)}\frac{q\left(\theta\middle|\theta^{*}\right)}{q\left(\theta^{*}\middle|\theta\right)}\end{aligned}$$

- In this MH algorithm, $X_{1:T}$ has been essentially integrated out.

Implementation Issues

- **Problem 1**: We do not know $p_\theta\left(y_{1:T}\right) = \int p_\theta\left(x_{1:T}, y_{1:T}\right) dx_{1:T}$ analytically.
- **Problem 2:** We do not know how to sample from $p_\theta\left(\left. x_{1:T}\right| y_{1:T}\right)$.
- **"Idea"**: Use SMC approximations of $p_\theta\left(\left. x_{1:T}\right| y_{1:T}\right)$ and $p_\theta\left(y_{1:T}\right)$.

Sequential Monte Carlo Approximation

- Given θ, SMC methods provide approximations $\widehat{p}_{\theta}\left(dx_{1:T} \middle| y_{1:T} \right)$ of $p_{\theta}\left(x_{1:T} \middle| y_{1:T} \right)$ and $\widehat{p}_{\theta}\left(y_{1:T} \right)$ of $p_{\theta}\left(y_{1:T} \right)$.
- These approximations degrade linearly with T instead of exponentially under some regularity assumptions.
- **Problem**: We cannot compute analytically the particle filter proposal $q_{\theta}\left(dx_{1:T} \middle| y_{1:T} \right) = \mathbb{E}\left[\widehat{p}_{\theta}\left(dx_{1:T} \middle| y_{1:T} \right) \right]$ as it involves an expectation w.r.t all the variables appearing in the particle algorithm...

"Idealized" Marginal MH Sampler

At iteration i

- Sample $\theta^{*} \sim q\left(\cdot|\,\theta\left(i-1\right)\right)$.
- Sample $X_{1:T}^{*} \sim p_{\theta^{*}}\left(\cdot|\,y_{1:T}\right)$.
- With probability

$$1 \wedge \frac{p_{\theta^{*}}\left(y_{1:T}\right)p\left(\theta^{*}\right)}{p_{\theta(i-1)}\left(y_{1:T}\right)p\left(\theta\left(i-1\right)\right)}\frac{q\left(\theta\left(i-1\right)|\,\theta^{*}\right)}{q\left(\theta^{*}|\,\theta\left(i-1\right)\right)}$$

set $\theta\left(i\right)=\theta^{*}$, $X_{1:T}\left(i\right)=X_{1:T}^{*}$ otherwise set $\theta\left(i\right)=\theta\left(i-1\right)$, $X_{1:T}\left(i\right)=X_{1:T}\left(i-1\right)$.

Particle Marginal MH Sampler

At iteration i

- Sample $\theta^* \sim q\left(\cdot \mid \theta\left(i-1\right)\right)$ and run an SMC algorithm to obtain $\widehat{p}_{\theta^*}\left(dx_{1:T} \mid y_{1:T}\right)$ and $\widehat{p}_{\theta^*}\left(y_{1:T}\right)$.
- Sample $X_{1:T}^* \sim \widehat{p}_{\theta^*}\left(\cdot \mid y_{1:T}\right)$.
- With probability

$$1 \wedge \frac{\widehat{p}_{\theta^*}\left(y_{1:T}\right) p\left(\theta^*\right)}{\widehat{p}_{\theta(i-1)}\left(y_{1:T}\right) p\left(\theta\left(i-1\right)\right)} \frac{q\left(\theta\left(i-1\right) \mid \theta^*\right)}{q\left(\theta^* \mid \theta\left(i-1\right)\right)}$$

set $\theta\left(i\right) = \theta^*$, $X_{1:T}\left(i\right) = X_{1:T}^*$ otherwise set $\theta\left(i\right) = \theta\left(i-1\right)$, $X_{1:T}\left(i\right) = X_{1:T}\left(i-1\right)$.

Validity of the Particle Marginal MH Sampler

- Assume that the 'idealized' marginal MH sampler is irreducible and aperiodic then, under very weak assumptions, the PMMH sampler generates a sequence $\{\theta(i), X_{1:T}(i)\}$ whose marginal distributions $\left\{\mathcal{L}^N(\theta(i), X_{1:T}(i) \in \cdot)\right\}$ satisfy for any $N \geq 1$

$$\left\|\mathcal{L}^N(\theta(i), X_{1:T}(i) \in \cdot) - p(\cdot | y_{1:T})\right\|_{\text{TV}} \rightarrow 0 \text{ as } i \rightarrow \infty \text{ .}$$

- Corollary of a more general result: the PMMH sampler is a standard MH sampler of target distribution $\tilde{\pi}^N$ and proposal $\widetilde{q}^N$ defined on an extended space associated to all the variables used to generate the proposal.

Explicit Structure of the Target Distribution

- Let first consider the case where $T = 1$.
- *Proposal distribution*

$$\widetilde{q}^N\left(\left(\theta^*, k, x_1^{1:N}\right)\middle|\,\theta\right) = q\left(\theta^*\middle|\,\theta\right)\ \prod_{m=1}^{N} \mu_{\theta^*}\left(x_1^m\right)\ w_1^k$$

- *Target distribution*

$$\tilde{\pi}^N\left(\theta, k, x_1^{1:N}\right) \propto \underbrace{\frac{1}{N}\sum_{m=1}^{N} g_\theta\left(y_1\middle|\,x_1^m\right)}_{\widehat{p}_\theta(y_1)}\ p\left(\theta\right)\ \prod_{m=1}^{N} \mu_\theta\left(x_1^m\right)\ w_1^k$$

- We have indeed

$$\frac{\tilde{\pi}\left(\theta^*, k, x_1^{1:N}\right)}{\widetilde{q}^N\left(\left(\theta^*, k, x_1^{1:N}\right)\middle|\,\theta\right)} = \frac{p\left(\theta^*\right)}{q\left(\theta^*\middle|\,\theta\right)}\frac{\widehat{p}_{\theta^*}\left(y_1\right)}{p_{\theta^*}\left(y_1\right)}$$

Explicit Structure of the Target Distribution

- As we have
$$\mathbb{E}\left(\widehat{p}_{\theta}\left(y_1\right)\right)=p_{\theta}\left(y_1\right)$$
then it follows that
$$\tilde{\pi}^N\left(\theta\right)=p\left(\left.\theta\right|y_1\right).$$
- However, we can actually rewrite the target as
$$\tilde{\pi}^N\left(\theta,k,x_1^{1:N}\right)=\frac{p\left(\left.\theta,x_1^k\right|y_1\right)}{N}\prod_{m=1;m\neq k}^{N}\mu_{\theta}\left(x_1\right).$$
- This shows that we are able to sample from $p\left(\left.\theta,x_1\right|y_1\right)$ and not only its marginal $p\left(\left.\theta\right|y_1\right)$.

Sampling from the Target Distribution

- To sample from this target distribution
  - Sample K from a uniform distribution on $\{1, ..., N\}$.
  - Sample $\left(\theta, X_1^K\right)$ from $p\left(\theta, x_1 \middle| y_1\right)$. (We do not know how to do this, this is why we use MCMC).
- Sample $X_1^m \sim \mu_\theta\left(\cdot\right)$ for $m \neq K$.

Ancestral Lines Generated by SMC

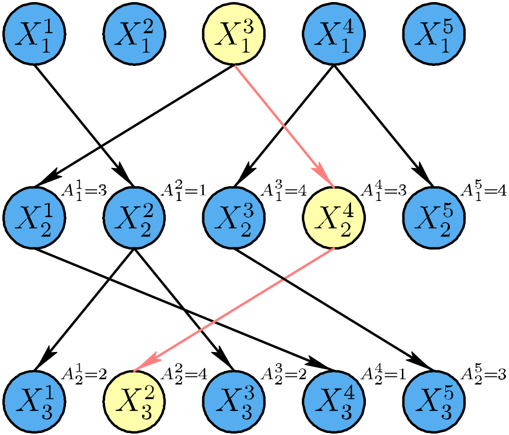


Figure: Ancestral lineages for $N = 5$ and $T = 3$. The lighter path is $X_{1:3}^2 = \left(X_1^3, X_2^4, X_3^2\right)$ and its ancestral lineage $B_{1:3}^2 = (3, 4, 2)$

Structure of the Proposal and Target Distributions

- *Proposal distribution*

$$q^N(\theta^*, k, \mathbf{x}_1, \ldots, \mathbf{x}_T, \mathbf{a}_1, \ldots, \mathbf{a}_{T-1} | \theta) \\ = q(\theta^* | \theta)\ \psi^{\theta^*}(\mathbf{x}_1, ..., \mathbf{x}_T, \mathbf{a}_1, \ldots, \mathbf{a}_{T-1})\ w_T^k$$

where

$$\psi^\theta(\mathbf{x}_1, \ldots, \mathbf{x}_T, \mathbf{a}_1, \ldots, \mathbf{a}_{T-1}) \\ = \left(\prod_{m=1}^{N} \mu_\theta(x_1^m)\right) \prod_{n=2}^{T} \left(r(\mathbf{a}_{n-1} | \mathbf{w}_{n-1}) \prod_{m=1}^{N} f_\theta(x_n^m | x_{n-1}^{a_{n-1}^m})\right)$$

- *Target distribution*

$$\tilde{\pi}^N(\theta, k, \mathbf{x}_1, \ldots, \mathbf{x}_T, \mathbf{a}_1, \ldots, \mathbf{a}_{T-1}) \\ \propto \widehat{p}_\theta(y_{1:T})\, p(\theta)\ \psi^\theta(\mathbf{x}_1, \ldots, \mathbf{x}_T, \mathbf{a}_1, \ldots, \mathbf{a}_{T-1})\ w_T^k$$

Explicit Structure of the Target Distribution

- The target can be rewritten as

$$
\begin{aligned}
&\tilde{\pi}^{N}\left(\theta, k, \mathbf{x}_1, \ldots, \mathbf{x}_T, \mathbf{a}_1, \ldots, \mathbf{a}_{T-1}\right) \\
&= \frac{p\left(\theta, x_{1:T}^{k} \middle| y_{1:T}\right)}{N^{T}} \frac{\psi^{\theta}\left(\mathbf{x}_1, \ldots, \mathbf{x}_T, \mathbf{a}_1, ..., \mathbf{a}_{T-1}\right)}{\mu_\theta(x_1^{b_1^k}) \prod_{n=2}^{T}\left(r(b_{n-1}^{k}|\mathbf{w}_{n-1}) f_\theta\left(x_n^{b_n^k} \middle| x_{n-1}^{b_{n-1}^k}\right)\right)},
\end{aligned}
$$

- To sample from this target distribution
  - Sample $\left(B_1^K, B_2^K, ..., B_{T-1}^K, K\right)$ from a uniform distribution on $\{1, ..., N\}^T$.
  - Sample θ and $X_{1:T}^{K} = (X_1^{B_1^K}, X_2^{B_2^K}, \ldots, X_{T-1}^{B_{T-1}^K}, X_T^K)$ from $p\left(\theta, x_{1:T} \middle| y_{1:T}\right)$. (We do not know how to do this, this is why we use MCMC).
- Run a conditional SMC algorithm compatible with $X_{1:T}^{K}$ and its ancestral lineage $\left(B_1^K, B_2^K, ..., B_{T-1}^K, K\right)$.

Conditional SMC

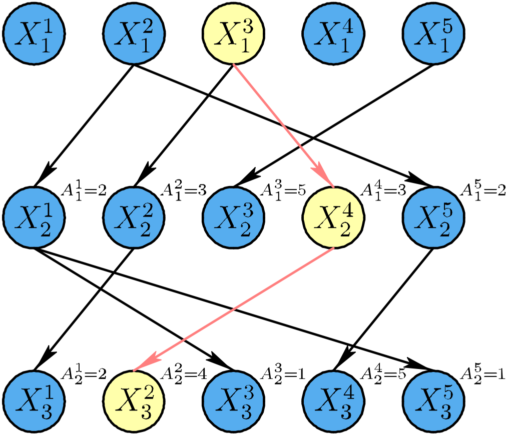


Figure: Example of $N-1=4$ ancestral lineages generated by a conditional SMC algorithm for $N=5$, $T=3$ conditional upon $X_{1:3}^2$ and $B_{1:3}^2$

Conditional SMC Algorithm

At time 1

- For $m \neq b_1^k$, sample $X_1^m \sim \mu_\theta(\cdot)$ and set $W_1^m \propto g_\theta(y_1 | X_1^m)$, $\sum_{m=1}^N W_1^m = 1$.
- Resample $N-1$ times from $\widehat{p}_\theta(dx_1 | y_1) = \sum_{m=1}^N W_1^m \delta_{X_1^m}(dx_1)$ to obtain $\left\{ \overline{X}_1^{-b_1^k} \right\}$ and set $\overline{X}_1^{b_1^k} = X_1^{b_1^k}$.

At time $n = 2, ..., T$

- For $m \neq b_n^k$, sample $X_n^m \sim f_\theta\left(\cdot | \overline{X}_{n-1}^m\right)$, set $X_{1:n}^m = \left(\overline{X}_{1:n-1}^m, X_n^m\right)$ and $W_n^m \propto g_\theta(y_n | X_n^m)$, $\sum_{m=1}^N W_n^m = 1$.
- Resample $N-1$ times from $\widehat{p}_\theta(dx_{1:n} | y_{1:n}) = \sum_{m=1}^N W_n^m \delta_{X_{1:n}^m}(dx_{1:n})$ to obtain $\left\{ \overline{X}_{1:n}^{-b_n^k} \right\}$ and set $\overline{X}_{1:n}^{b_n^k} = X_{1:n}^{b_n^k}$.

At time $n = T$

- Sample $X_{1:T} \sim \widehat{p}_\theta(dx_{1:T} | y_{1:T})$.

“Idealized” Gibbs Sampler

- To sample from $p\left(\theta, x_{1:T}\middle|\, y_{1:T}\right)$, an MCMC strategy consists of using the following block Gibbs sampler.

At iteration i

- Sample $X_{1:T}\left(i\right) \sim p_{\theta(i-1)}\left(\cdot\middle|\, y_{1:T}\right)$.
- Sample $\theta\left(i\right) \sim p\left(\cdot\middle|\, y_{1:T}, X_{1:T}\left(i\right)\right)$.

- **Problem**: We do not know how to sample from $p_{\theta}\left(x_{1:T}\middle|\, y_{1:T}\right)$.
- Naive particle approximation where $X_{1:T}\left(i\right) \sim \widehat{p}\left(\cdot|y_{1:T}, \theta\left(i\right)\right)$ is substituted to $X_{1:T}\left(i\right) \sim p\left(\cdot|y_{1:T}, \theta\left(i\right)\right)$ is obviously incorrect.

Particle Gibbs Sampler

At iteration i

- Sample $\theta\left(i\right)\sim p\left(\cdot|y_{1:T},X_{1:T}\left(i-1\right)\right)$.
- Run a conditional SMC algorithm for $\theta\left(i\right)$ consistent with $X_{1:T}\left(i-1\right)$ and its ancestral lineage.
- Sample $X_{1:T}\left(i\right)\sim\widehat{p}\left(\cdot|y_{1:T},\theta\left(i\right)\right)$ from the resulting approximation (hence its ancestral lineage too).

- **Proposition**. Assume that the 'ideal' Gibbs sampler is irreducible and aperiodic then under very weak assumptions the particle Gibbs sampler generates a sequence $\left\{\theta\left(i\right),X_{1:T}\left(i\right)\right\}$ such that for any $N\geq 2$

$$\left\|\mathcal{L}\left(\left(\theta\left(i\right),X_{1:T}\left(i\right)\right)\in\cdot\right)-p(\cdot|\,y_{1:T})\right\|\rightarrow 0\text{ as }i\rightarrow\infty.$$

Nonlinear State-Space Model

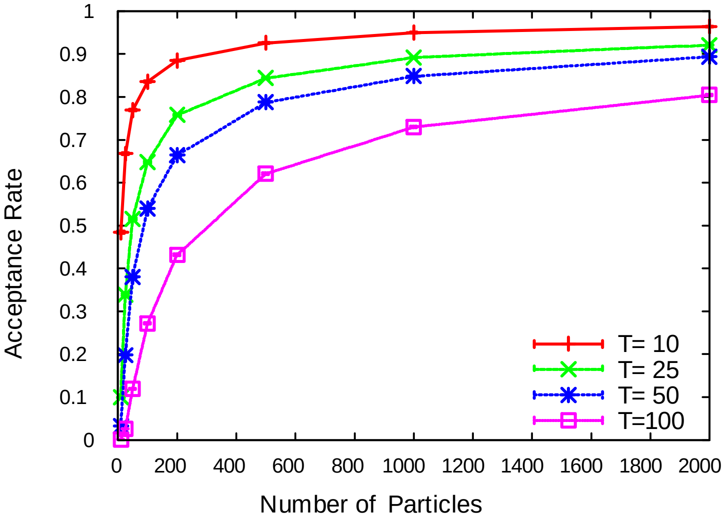
- Consider the following model

$$\begin{aligned} X_n &= \frac{1}{2}X_{n-1} + 25\frac{X_{n-1}}{1+X_{n-1}^2} + 8\cos 1.2n + V_n, \\ Y_n &= \frac{X_n^2}{20} + W_n \end{aligned}$$

  where $V_n \sim \mathcal{N}\left(0, \sigma_v^2\right)$, $W_n \sim \mathcal{N}\left(0, \sigma_w^2\right)$ and $X_1 \sim \mathcal{N}\left(0, 5^2\right)$.

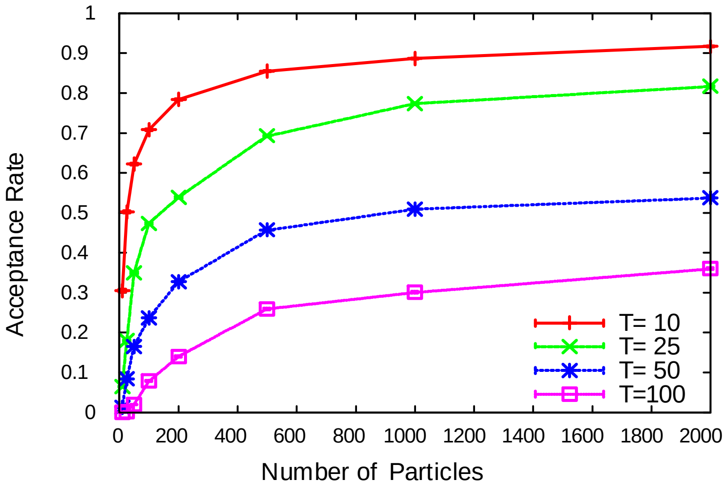
- Use the prior for $\{X_n\}$ as proposal distribution.
- For a fixed θ, we evaluate the expected acceptance probability as a function of N.

Average Acceptance Probability



Average acceptance probability when $\sigma_v^2 = \sigma_w^2 = 10$

Average Acceptance Probability



Average acceptance probability when $\sigma_v^2 = 10$, $\sigma_w^2 = 1$

Inference for Stochastic Kinetic Models

- Two species X_t^1 (prey) and X_t^2 (predator)

$$\begin{aligned}
&\Pr\left(X_{t+dt}^1 = x_t^1 + 1, X_{t+dt}^2 = x_t^2 \,\middle|\, x_t^1, x_t^2\right) = \alpha\, x_t^1\, dt + o\left(dt\right), \\
&\Pr\left(X_{t+dt}^1 = x_t^1 - 1, X_{t+dt}^2 = x_t^2 + 1 \,\middle|\, x_t^1, x_t^2\right) = \beta\, x_t^1\, x_t^2\, dt + o\left(dt\right), \\
&\Pr\left(X_{t+dt}^1 = x_t^1, X_{t+dt}^2 = x_t^2 - 1 \,\middle|\, x_t^1, x_t^2\right) = \gamma\, x_t^2\, dt + o\left(dt\right),
\end{aligned}$$

observed at discrete times

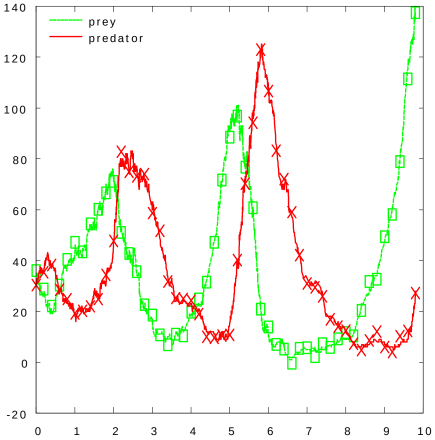
$$Y_n = X_{n\Delta}^1 + W_n \text{ with } W_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2\right).$$

- We are interested in the kinetic rate constants $\theta = (\alpha, \beta, \gamma)$ a priori distributed as (Boys et al., 2008; Kunsch, 2011)

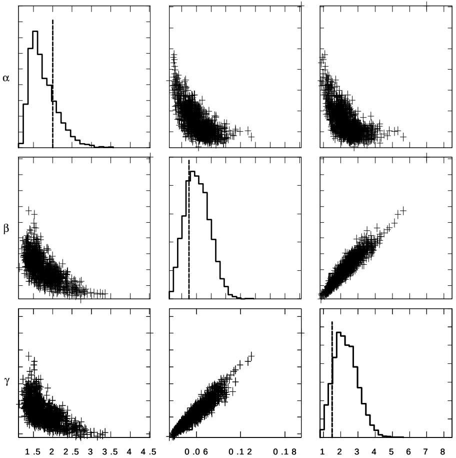
$$\alpha \sim \mathcal{G}(1, 10), \quad \beta \sim \mathcal{G}(1, 0.25), \quad \gamma \sim \mathcal{G}(1, 7.5).$$

- MCMC methods require reversible jumps, Particle MCMC requires only forward simulation.

Experimental Results

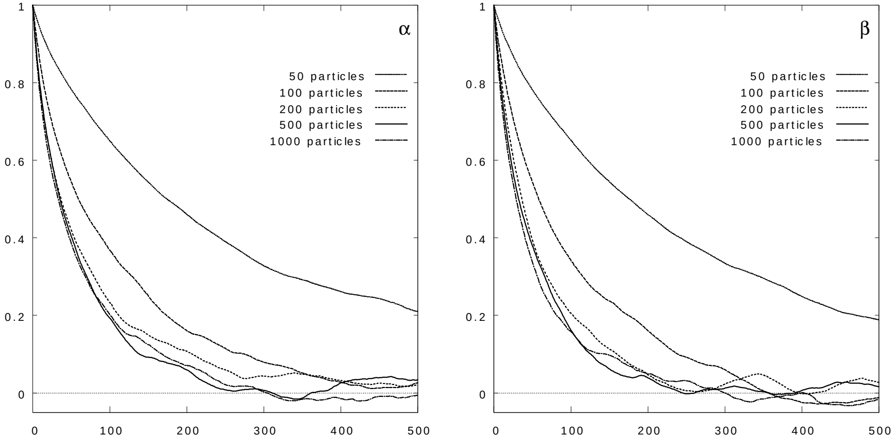


Simulated data



Posterior distributions

Autocorrelation Functions



Autocorrelation of α (left) and β (right) for the PMMH sampler for various N.

Discussion

- PMCMC methods allow us to design 'good' high dimensional proposals based only on low dimensional (and potentially unsophisticated) proposals.
- PMCMC allow us to perform Bayesian inference for dynamic models for which only forward simulation is possible.
- "Computationally brutal" but several implementations on GPU already available and applications in ecology, econometrics (Flury & Shephard, *Econometrics Review*, 2011), biochemical systems, epidemiology etc.
- More precise quantitative convergence results need to be established.